CUARTA PARTE

EL ANÁLISIS Y LA PRESENTACIÓN DE LA INFORMACIÓN

9

EL ANÁLISIS DE LOS DATOS

Una vez que ha concluido la recogida de información, comienza la fase de *análisis e interpretación de los datos*. Ésta es una fase compleja e igualmente crucial en el proceso de investigación. Si bien, en los últimos años se ha visto beneficiada por los grandes avances producidos en el campo de la informática. En concreto, la expansión de paquetes estadísticos informáticos (adaptados a ordenadores personales), que cubren el amplio espectro analítico existente (tanto *cuantitativo* como *cualitativo).* La implantación de estos programas ha adquirido tal relieve que, en la actualidad, sería inconcebible un *análisis de datos* ajeno al procesamiento informático.

En las páginas siguientes se ofrece una panorámica de las *técnicas cuantitativas de análisis de datos*. La notable pluralidad de éstas, unido a condicionantes de espacio, limitan su exposición a una mera descripción de rasgos distintivos de cada una de ellas. Para un mayor conocimiento y profundización, se remite a la consulta de bibliografía especializada en *análisis de datos* (véase la selección de textos que figuran al final del capítulo, y en la bibliografía general al término del libro).

9.1. Los preliminares del análisis de datos

"El paso más importante antes de comenzar a diseñar un proceso de entrada de datos es determinar qué programas se utilizarán para analizar los datos y convenciones concretas respecto a los formatos de ficheros y de *missing data,* que pueden manejarse para esos programas" (Fowler, 1988: 128). De ahí que la exposición de las *técnicas estadísticas de análisis* comience por sus *preliminares* esenciales: la creación de los *ficheros de datos,* junto a la depuración de la información (con especial referencia a la "no-respuesta").

9.1.1. La creación de los ficheros de datos

Cada programa informático comprende instrucciones concretas para la creación de *ficheros de datos* propios y *ficheros de definición de los datos*. No obstante, la mayoría de estos programas permiten el acceso a ficheros elaborados por otros programas.

Primero, se confecciona un *fichero de datos* para la introducción de la información obtenida durante la investigación. Los *ficheros de datos* presentan formas diversas, dependiendo del programa informático que se maneje. La más común es el fichero de datos en formato *ASCII (American Standard Code for Information Interchange)*. Este se compone de filas y columnas de números. Cada *fila* representa los datos recabados de cada sujeto o caso observado. En las *columnas,* en cambio, figuran los valores de las variables correspondientes a cada caso (véase el Cuadro 9.1). Este tipo de formato se precisa para ejecutar la mayoría de los *paquetes estadísticos genéricos* (como el SPSS, el BMDP o el SAS).

CUADRO 9.1. Extracto de un fichero de datos.

```
0012116114500214020325402100110119342124164211021110053121232112005020 13
0023129113400102920163322210121013211222162212102111201432236125202020 41
0038116124300243320163411106111143003111111200002211302711244132113111 22
0049112911500214020325302001432214002221221110031321222320402071100021 10
0052119114500212120111411109233221130022211211211010011044223426140011 03
0063118114607312420145422109234112300121111211202002202432232211300310 012
0076117114400212120175411110423112300122112222000010024322362110400532 12
```

Los *paquetes estadísticos genéricos* también pueden, sin embargo, leer *ficheros de datos* creados por *bases de datos* (como el DBASE) u *hojas de cálculo* (como LOTUS). Estos ficheros incluyen nombres para las variables, además de información relativa a la amplitud de la variable y la *codificación* de los valores de "no-respuesta" *(missing data).*

Cuando los datos se han registrado en *formato ASCII,* hay que elaborar también un *fichero de definición de datos*. Este comprende instrucciones precisas para la lectura de los datos que componen el *fichero de datos*. Estas instrucciones incluyen información primordial sobre las variables analizadas en la investigación: *columnas* donde se hallan ubicadas en el *fichero de datos;* las *etiquetas* dadas a las variables y a sus atributos; así como la especificación de cómo se *codifican* los valores *"missing"* (o sin respuesta), para su tratamiento en el análisis de los datos.

En el Cuadro 9.2 figura un extracto de un *fichero de definición de datos* en SPSS. Comienza con el nombre del fichero donde se han grabado los datos. Prosigue con la relación de variables y el número de las columnas donde éstas se encuentran registradas. Después, se asignan *etiquetas* a las variables que lo precisen, al igual que a sus valores. En el SPSS, como en los demás programas informáticos, se dan al usuario instrucciones

para la correcta elaboración de estos ficheros. También se expone cómo transportar y convertir ficheros de otros programas, su lectura y demás especificaciones necesarias para el análisis de los datos.

CUADRO 9.2. Extracto de un fichero de definición de datos en SPSS.

```
Data list file= 'menores.dat'.
/id 1-7 anio 8 sexo 9 edad 10 nacional 11 etnia 12 anorma 13 alcohol 14
heroina 15 cannabis 16 estudio1 to estudio2 17-20 desescol 21 profepa 22
situacpa 23 apreviv 24 ambiente 25 fuga 26 compania 27.
Variable labels desescol 'desescolaridad'
/profepa 'profesion del padre'
/situacpa 'situación empleo padre'
/apreviv 'apreciación vivienda'.
Value labels sexo 1 'varon' 2 'mujer'
/nacional 1 'espanol' 2 'extranjero'
/estudio1 to estudio2 01 'primero EGB' 02 'segundo EGB' 03 'tercero EGB'
04 'cuarto EGB' 05 'quinto EGB' 06 'sexto EGB' 07 'septimo EGB'
08 'octavo EGB' 09 'FP' 10 'BUP' 11 'compensatoria' 12 'otros'
13 'analfabeto'.
Missing values anio to compania (0).
Frecuencies anio to compania /statistics=all /hbar.
Save outfile='menores.sys'.
```

Aparte de estos ficheros, pueden crearse *subficheros* específicos con objeto de facilitar los análisis. Especialmente, cuando el tamaño de la muestra es elevado y el *fichero de datos* originario adquiere un gran volumen.

Estos *subficheros* incluyen una *muestra de variables,* restringida a aquellas que sean de interés para la ejecución de análisis concretos. También, pueden representar una submuestra aleatoria de los casos observados en la investigación. En este último caso, la finalidad no sería tanto el ahorro de tiempo en el análisis de la información, sino la *validación* de los resultados estadísticos obtenidos de la otra submuestra de la muestra global. Así, por *ejemplo,* para la *validación* de *análisis multivariables* (como el de *regresión),* se recomienda la división de la *muestra* en dos *submuestras:* una, para la consecución del modelo; y la otra, para su *validación.*

9.1.2. La depuración de la información

A la creación de ficheros sigue la *depuración de los datos*, como antesala del análisis. El investigador ha de identificar posibles *errores* cometidos en la *grabación de los datos.* A tal fin resulta conveniente solicitar al programa la relación de *frecuencias* de todas las *variables* introducidas en el *fichero de datos* (*"frequencies"* en SPSS, o *"univariate"*

en SAS, por ejemplo). Esta relación incluye un listado con todos los *valores* de cada *variable*, la *frecuencia (absoluta y relativa)* de cada uno de ellos, los casos sin respuesta *(missing data), estadísticos univariados* y las *representaciones gráficas* que se soliciten.

De esta relación, se observará si existen anomalías en los *valores* de las *variables* codificadas. Más concretamente, si alguna de ellas incluye valores ajenos al recorrido o *rango* definido de la *variable*.

Por *ejemplo*, si en la variable sexo, que se ha codificado con sólo dos opciones de respuesta (1 'varon' 2 'mujer'), aparecen casos con valores superiores a dos (3, 4, 5 u otro), éstos corresponderían a sujetos que han sido erróneamente codificados o grabados en el ordenador. Por lo que, habría que proceder a su comprobación y corrección posterior.

Algunos programas informáticos (como el SAS o el SPSS) proporcionan, además, especificaciones dirigidas a la *depuración de los datos* (tanto durante su introducción, como una vez concluida ésta).

Si se observan incongruencias en los *valores* de las *variables* (como la anteriormente ejemplificada), ha de procederse a su localización y corrección. Los *errores* pueden deberse a fallos en la introducción de los datos en el ordenador, pero no siempre. También pueden ser ocasionados por deficiencias en la recogida de información.

En el primer caso, la corrección resulta más viable e inmediata: se revisan los *cuestionarios* (u otro instrumento de recogida de datos que se haya empleado), hasta localizar los casos en que se han grabado mal los *valores* de las *variables;* posteriormente, se introducirían los *códigos* correctos de las *variables* correspondientes.

Pero, si los *errores* se deben a un mal registro de la información en el instrumento de medición, las posibilidades de corrección se restringen. La dificultad de contactar de nuevo con la fuente de información (las *unidades muestrales)* lleva, no a la transformación de *códigos numéricos,* sino a la eliminación de aquellos casos con datos incorrectos o inconsistentes. Éstos se sumarían a aquellos que originariamente no proporcionaron información alguna. De esta forma, se incrementaría el volumen de los llamados *"missing values"* (o *valores* con los que se codifican las respuestas en blanco o incorrectas, y que se dan por *perdidos).*

Cuando unos casos específicos presentan muchos *"missing values"* en la mayoría de las *variables,* suele decidirse su exclusión del *fichero de datos* (salvo que al investigador le interese el análisis y descripción de estos casos). Igualmente, si de alguna variable se obtiene escasa información (teniendo un elevado porcentaje de valores *missing),* también suele optarse por su exclusión para el resto de los análisis.

En general, antes de proceder al *análisis de los datos,* el investigador evalúa los *porcentajes de respuesta* (para cada *variable),* y los *"outliers"* registrados en la *matriz de datos.*

Por "*outliers*" se entiende cualquier observación o caso que muestre inconsistencia con la serie global de datos. Su identificación requiere la realización de *análisis univariables* (para cada una de las *variables),* tanto numéricos como gráficos. Además, la mayoría de los programas informáticos incluyen instrucciones específicas para la detección de *"outliers".*

En cuanto a la *"no-respuesta",* su evaluación resulta igualmente exigida. Diversos autores, como Bourque y Clark (1994), recomiendan la *comparación* de las caracte-

rísticas demográficas de la *muestra* con las correspondientes a la *población* de la que ésta procede. Para ello se emplea el *Censo de Población,* u otra fuente de datos estadísticos o estudio, que describa al conjunto de la *población.*

Si de este análisis se dedujese la no *representatividad* de la *muestra,* el investigador deberá establecer la magnitud de las diferencias entre la *población* y la *muestra.* Esto es importante para la delimitación de las posibilidades de *inferencia* de las *estimaciones muestrales.* En palabras de Arber (1993: 71):

> "La capacidad para realizar inferencias de una muestra a una población se basa en el supuesto de que la muestra lograda no esté sesgada por la *no-respuesta.* En la medida en que aquellos que no responden difieran de forma significativa de aquellos que sí responden, el investigador tiene una muestra sesgada."

9.2. El análisis estadístico univariable

En el diseño de la investigación ya se prevén los análisis a realizar con la información reunida en el desarrollo de la investigación. Aunque ha de matizarse que el *proyecto de análisis* no es inmutable. Depende, en gran parte, de la *cantidad* y *calidad* de los datos que se recaben. De ahí la importancia que adquieren, en cualquier indagación, los *análisis exploratorios,* como paso exigido y previo a la decisión de qué técnica analítica *(bivariable* y *multivariable)* se va a aplicar.

9.2.1. La distribución de frecuencias

En la *exploración de los datos,* primero se procede a un análisis exhaustivo de cada *variable* incluida en la *matriz de datos (análisis univariable).* Para cada una de las *variables* se calcula su *distribución o tabla de frecuencias.* La *tabla de frecuencias* –como puede verse en el Cuadro 9.3– incluye los distintos *valores* que presenta la *variable* (distribuidos en *clases* o *categorías),* acompañados por su *frecuencia* (es decir, el número de veces en que aparecen).

En la primera columna (encabezada por el rótulo *"value label")* figuran los distintos atributos que componen la *variable.* La siguiente columna *("value")* muestra el *valor* dado a cada *atributo.* En la tercera columna (la denominada *"frecuency")* se hallan las *frecuencias absolutas;* o sea, el número de casos (de la *muestra)* que comparten cada uno de los *valores* de la *variable.*

Para conocer la importancia de cada *valor,* y a efectos comparativos, se obtienen las *frecuencias relativas o porcentuales,* que representan cada *valor* en el conjunto de la *muestra* (columnas 4.ª y 5.ª). Primero, se calculan los *porcentajes* para toda la *muestra;* segundo, exclusivamente para aquellos casos que han proporcionado información al respecto (columna de *"valid percent"* o porcentaje válido). En esta columna no se consideran, por tanto, los *"missing values".*

Este desglose de casos, en función de si aportan o no información, permite conocer la proporción de *"no-respuesta"* de cada *variable.* Este conocimiento adquiere especial relevancia para posteriores análisis.

Por último, se calculan las *frecuencias relativas acumuladas ("cum percent")*, a partir de las *frecuencias* contenidas en la columna 5.ª. Estas *frecuencias acumuladas* denotan la proporción de casos (válidos) que se encuentran por debajo, o por encima, de un determinado *valor* de la *variable*.

Cuando la *variable* está medida a nivel de *intervalo*, se aconseja la previa *agrupación* de los *valores*. Ello facilitará su presentación en una *tabla de frecuencias* de menores dimensiones. A este respecto, algunos autores –como Bryman y Cramer (1995)– sugieren que el número de *categorías* diferenciadas esté comprendido entre 6 y 20. Argumentan que menos de 6 y más de 20 categorías pueden distorcionar la *forma* de la distribución de la variable.

Los *estadísticos univariables* que figuran en el Cuadro 9.3 se comentan en la sección 9.2.3, dedicada a su exposición.

CUADRO 9.3. Ejemplo de una tabla de frecuencias para la variable "estudio" en SPSS.

Value label	Value	Frecuency	Percent	Valid Percent	Cum Percent
Primero EGB	1	8	1.1	1.1	1.1
Segundo EGB	2	8	1.1	1.1	2.2
Tercero EGB	3	31	4.3	4.3	6.6
Cuarto EGB	4	59	8.3	8.3	14.8
Quinto EGB	5	119	16.6	16.7	31.5
Sexto EGB	6	152	21.3	21.3	52.8
Septimo EGB	7	177	24.8	24.8	77.6
Octavo EGB	8	72	10.1	10.1	87.7
FP	9	27	3.8	3.8	91.5
BUP	10	7	1.0	1.0	92.4
Compensatoria	11	1	.1	.1	92.6
Analfabeto	13	53	7.4	7.4	100.0
.	0	1	.1	MISSING	
	TOTAL	715	100.0	100.0	

Primero EGB 8
Segundo EGB 8
Tercero EGB 31
Cuarto EGB 59
Quinto EGB 119
Sexto EGB 152
Septimo EGB 177
Octavo EGB 72
FP 27
BUP 7
Compensatoria 1
Analfabeto 53

Mean	6.566	Std Err	.091	Median	6.000
Mode	7.000	Std Dev	2.420	Variance	5.856
Kurtosis	1.644	S E Kurt	.183	Skewness	1.000
S E Skew	.091	Range	12.000	Minimum	1.000
Maximum	13.000	Sum	4688.000		
Valid Cases	714	Missing Cases	1		

9.2.2. Representaciones gráficas

La información contenida en una *tabla de frecuencias* también puede representarse de forma gráfica. Ello ayuda a la visualización global de la concentración, o dispersión, de los datos en la variable considerada.

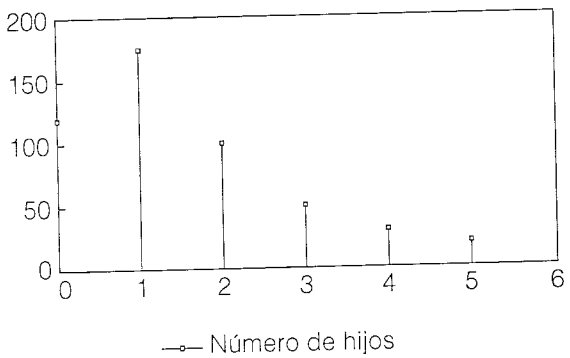
Dos de los *gráficos* habituales en la representación de *frecuencias* son el *diagrama de barras* y el *histograma*. A ellos se suman otros también usuales en el *análisis exploratorio*, como el diagrama de *"tronco y hoja"* y la *"caja"*; o el *polígono*, las *ojivas*, y el *gráfico de sectores*, entre la amplia variedad gráfica existente.

- *Diagrama de barras*

Consiste en una serie de "barras" (una para cada categoría de la variable), cuyas longitudes expresan las *frecuencias* de cada *atributo* de la *variable*.

EJEMPLO DE DIAGRAMA DE BARRAS

Número de hijos	Frecuencia absoluta
0	120
1	170
2	90
3	50
4	25
5	15
	470



- *Histograma*

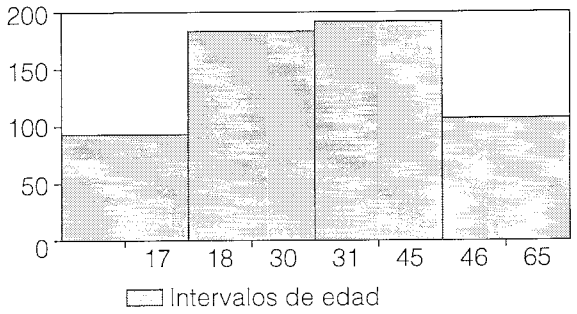
Un gráfico de contenido similar al anterior, aunque más adecuado a la representación de variables *métricas* (de *intervalo* o de *razón*).

En lugar de barras, el gráfico se compone de *rectángulos* unidos. Razón por la que se exige que la variable esté agrupada en *intervalos*. Estos forman las bases de cada uno de los rectángulos, que están delimitados por los *límites (reales)* de los respectivos *intervalos*. En cambio, la altura de los rectángulos es proporcional a la *frecuencia (absoluta* o *relativa)* de cada *intervalo*.

La suma total de las áreas de los rectángulos será igual a 1 (dado que la suma de todas las proporciones es la unidad).

EJEMPLO DE HISTOGRAMA

Edad	Frecuencia absoluta
Menos de 18	90
18 – 30	170
31 – 45	185
46 – 65	115
	560



• *El "Tronco y las Hojas" ("Stem-and-Leaf")*

Constituye un gráfico parecido al *histograma,* pero integrado por los *dígitos* de los *valores* de las *variables*. Los *dígitos* se dividen entre dos. Los situados a la izquierda del punto (el *tronco)* figuran ordenados verticalmente, en orden creciente (de arriba a abajo). Por el contrario, los *dígitos* a la derecha del punto (las *hojas),* se disponen horizontalmente, aunque también en sentido creciente (de menor a mayor).

El *dígito* a la izquierda (la *columna)* que comprenda más *valores* a la derecha será aquél en el que se agrupen un mayor número de casos en la distribución. Por esta y otras razones, este gráfico suele tomarse como referente de las medidas de *tendencia central* de una distribución de frecuencias (a las que se hará referencia en el apartado 9.2.3).

EJEMPLO DEL GRÁFICO "EL TRONCO Y LAS HOJAS"

En el siguiente gráfico, puede observarse que la *fila* tercera representa los valores de mayor frecuencia en la distribución. En concreto, los *valores* 36, 36, 37, 37, 37, 38, 38, 39, 39, 39.

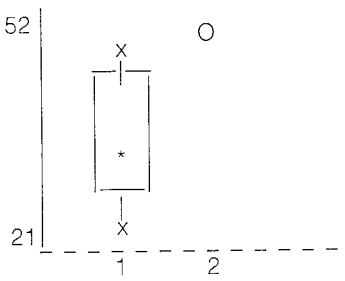
```
2 • 1 2 3 4
3 • 0 1 2 2 3 3 5
3 • 6 6 7 7 7 8 8 9 9
4 • 1 1 1 5 6 7 7
5 • 2 2
```

• *La "Caja" ("Box-plot o Box-and-Whisker")*

Ofrece una visión global de la distribución, más sintética que la anterior. En ella la *variable* se representa de forma que el 50% de los casos queden comprendidos en el interior de la "caja". En los extremos se sitúan, respectivamente, el 25% superior e inferior.

Con un asterisco se señala la *mediana;* con una "X" los valores máximos y mínimos; y con una "O", los *"outliers"*. De esta manera se proporciona (gráficamente) información referente a la *mediana,* el *primer cuartil* (el 25% de los casos iniciales), el *tercer cuartil* (el 25% finales), y el *recorrido intercuartílico* (el 50% de los casos centrales) de la distribución de frecuencias. Ello exige que el nivel de medición mínimo de la variable sea el *ordinal*. En caso contrario, no podría estimarse el valor de la *mediana*, ni de ningún estadístico que precise de la ordenación de los *valores* de la variable, en un sentido creciente o decreciente (los *cuantiles).*

EJEMPLO DEL GRÁFICO LA "CAJA"



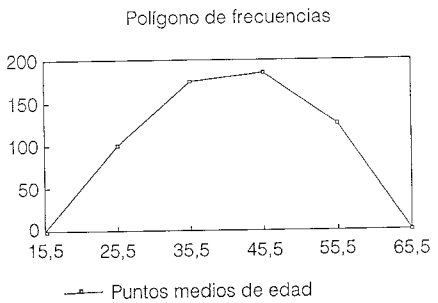
• *Polígono*

Un gráfico lineal que se traza sobre los *puntos medios* de cada *intervalo,* a una altura proporcional a su frecuencia *(absoluta* o *relativa).* Por lo que, de su visualización también se extraerán los *valores* con mayor (y menor) *frecuencia* en una distribución.

Los *puntos medios* se calculan sumando los *límites* de los *intervalos,* y dividiéndolo entre 2. De esta forma se obtiene el *valor* que representará al *intervalo* a todos los efectos. Por ejemplo, (21 + 30)/2 = 25,5.

EJEMPLO DE POLÍGONO DE FRECUENCIAS

Edad	Frecuencia absoluta	Punto medio
21 – 30	90	25,5
31 – 40	170	35,5
41 – 50	185	45,5
51 – 60	115	55,5
	560	



- *Ojivas*

Polígonos de frecuencias acumuladas que muestran la *frecuencia* de casos por encima, o por debajo, de un determinado *valor* de la distribución.

La *ojiva* será *"menor que"*, si se consideran los casos que hay por debajo de un *valor*. Por el contrario, será *"mayor que"*, cuando se representan los casos que comparten un *valor* superior de la distribución.

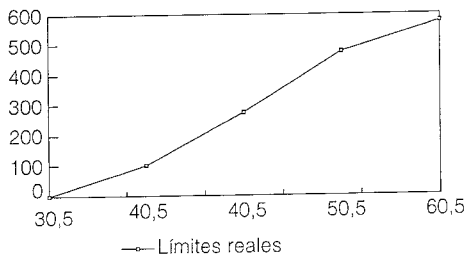
En el eje horizontal (de *abscisa)* se sitúan los *límites reales* de los *intervalos* de la *variable,* mientras que en el eje vertical (de *ordenada)* se disponen las *frecuencias acumuladas (absolutas* o *relativas)* de cada *intervalo.*

Para el cálculo de las *frecuencias acumuladas* se tiene en cuenta si interesa conocer el número (o la proporción) de casos que hay por debajo *(ojiva "menor que"),* o por encima *(ojiva "mayor que")* de un *valor* específico de la distribución.

EJEMPLO DE OJIVAS "MENOR QUE" Y "MAYOR QUE"

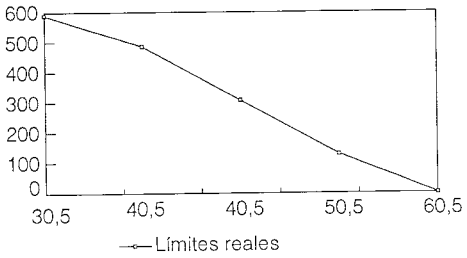
Ojiva "menor que"

Edad	Frecuencia absoluta	Frecuencia acumulada
21 – 30	90	90
31 – 40	170	260
41 – 50	185	445
51 – 60	115	560
	560	



Ojiva "mayor que"

Edad	Frecuencia absoluta	Frecuencia acumulada "más de"
21 – 30	90	560
31 – 40	170	470
41 – 50	185	300
51 – 60	115	115
	560	



• *Gráfico de sectores (o en forma de "tarta")*

A diferencia de los gráficos anteriores, éste se representa mediante un círculo, dividido en "sectores", cuyos *ángulos* indican el porcentaje de casos que comparten cada atributo de la variable.

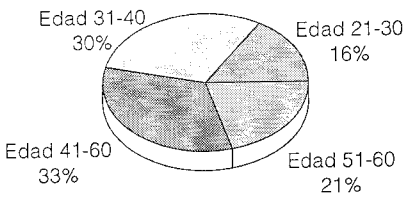
Para su obtención manual, se multiplican las *frecuencias relativas* (porcentual) de los *intervalos* (o valores) de la variable por 3,6 (que son los *grados* que corresponden a cada unidad porcentual: 360/100). De esta forma se transforman los *porcentajes* en *grados,* lo que posibilita su representación en una circunsferencia.

No obstante, el empleo de programas gráficos (como el Harvard Graphics) hace innecesaria la realización de estos cálculos. Basta con conocer la *frecuencia* de cada *valor* de la distribución.

Los "sectores" del gráfico suelen representarse con colores o trazas diferentes, que proporcionen un mayor contraste entre ellos.

EJEMPLO DE GRÁFICO DE SECTORES

Edad	Frecuencia absoluta	Frecuencia relativa (%)	Grados
21 – 30	90	16	58
31 – 40	170	30	108
41 – 50	185	33	119
51 – 60	115	21	75
	560		360



Para un conocimiento más amplio de éstas y demás representaciones gráficas comunes en la investigación social, se remite a textos específicos sobre *Gráficos* (como el de Alaminos (1993), y otros citados en la bibliografía).

9.2.3. Estadísticos univariables

Aparte de los *gráficos,* en la *exploración univariable* también se emplean estadísticos para medir, de manera precisa, la distribución de los valores de una variable (véase Cuadro 9.3). Su uso dependerá, sobre todo, del nivel de *medición* de la *variable.*

Los *estadísticos univariables* se agrupan en tres grandes apartados:

a) Medidas de tendencia central.
b) Medidas de dispersión.
c) Medidas de forma de la distribución.

• *Medidas de tendencia central*

Describen cómo se agrupan los *atributos* de una *variable* alrededor de un "valor típico" de la distribución. Por lo que, proporcionan una síntesis de la información contenida en la distribución.

Las medidas de *tendencia central* más empleadas en la investigación social son las siguientes: la *media,* la *mediana* y la *moda.*

1) La *media* es la medida más representativa, siempre y cuando la variable sea cuantitativa (de *intervalo* o de *razón).* Su cálculo precisa de la participación de todos los valores de la distribución. Cada uno de ellos se multiplica por sus respectivas *frecuencias absolutas.* Después, se suman todos los productos, y el resultado se divide por el número total de casos. De esta forma se obtiene el *promedio* de los valores de la distribución, que es como se define la *media aritmética.* El *inconveniente* fundamental de este estadístico es que se ve afectado por *valores* muy extremos en una distribución. Cuando la variable incluye *valores* muy dispares (o extremos), la *media* pierde representatividad a favor de la *mediana,* como *medida de tendencia central.*
2) La *mediana* es el valor que divide a la distribución en dos partes iguales. Se sitúa en el medio de la distribución. Lo que permite el conocimiento de los *valores* de mayor representación en la *muestra.* Su cálculo precisa que la variable sea, al menos, ordinal.
3) La *moda* denota el valor de mayor frecuencia en una distribución; aquel que más casos comparten. La distribución puede ser *unimodal* (una sola *moda),* pero también *bimodal* (dos *modas),* o *multimodal* (más de dos *modas).* Esto dificulta su interpretación, en caso de coexistir varias *modas.* A este *inconveniente* primordial se suma otro importante: en su cálculo no intervienen todos los *valores* de la distribución. Si bien, su principal *ventaja* es su universalidad. Pue-

de estimarse para cualquier tipo de variable, ya que el *nivel de medición* mínimo exigido es el *nominal.*

Además de estos estadísticos, existen los *cuantiles* como *medidas de tendencia no central.* Éstos representan *valores* que dividen a la distribución en partes iguales. Los *cuartiles* la dividen en cuatro partes iguales (cada una de ellas incluye al 25% de los *valores);* los *deciles* en diez partes; y los *percentiles* en cien partes. Su cálculo se asemeja al de la *mediana.* De hecho, el *cuartil* dos, el *decil* cinco, y el *percentil* cincuenta, expresan el *valor mediano.*

• *Medidas de dispersión*

Al conocimiento de los *valores centrales* de la distribución le sigue la medición de su representatividad: la mayor o menor *variabilidad* existente en torno a la *media* o la *mediana* de la distribución.

Las medidas de *dispersión absolutas* más comunes son el *rango* (o recorrido), la *desviación típica,* y la *varianza.*

1) El *rango o recorrido* expresa el número de valores incluidos en la distribución. Estos se obtienen de la diferencia entre el valor superior y el inferior. Su comprensión es sencilla, aunque presenta, en su contra, un *inconveniente* importante: es susceptible a la distorción proporcionada por valores muy extremos en una distribución. Ello se debe a la única consideración del primer y último valor de la distribución.

   Para solventar esta deficiencia, suelen aplicarse otras medidas de *rango,* que emplean un mayor volumen de información. Si bien, éstas exigen que la variable sea, al menos, *ordinal.* Se trata de los *recorridos: intercuartílico* (la diferencia entre el tercer *cuartil* y el primero; por lo que incluye al 50% de los valores centrales de la distribución), *semiintercuartílico* (el 25% de los valores centrales), *interpercentílico* (el 80% de los valores centrales, al ser la diferencia entre el *percentil* 90 y el 10), y el *semiinterpercentílico* (el 40%).
2) La *desviación típica* es el promedio de la desviación de los casos con respecto a la *media.* Como indicador de heterogeneidad (o de dispersión de los valores de una distribución), su estimación se exige siempre que se calcule la media, porque ayudará a la interpretación de su representatividad en la distribución.

   Al igual que la *media,* su valor viene expresado en la unidad de medición de la *variable,* y únicamente puede calcularse cuando la variable es *cuantitativa.*
3) La *varianza* constituye otra medida de heterogeneidad de una distribución. Se define como el cuadrado de la *desviación típica.* Su valor expresa el grado de heterogeneidad de una población respecto a la *variable* medida, siendo sus características similares a la *desviación.*

A partir de la *desviación*, puede calcularse una medida de *dispersión relativa* que ayudará en su interpretación. Se trata del *coeficiente de variabilidad de Pearson*. Una medida estandarizada que se obtiene del cociente entre el valor de la *desviación* y la *media aritmética*. Su valor se expresa en porcentajes, siendo de utilidad en la comparación de la homogeneidad de dos o más grupos respecto a una o más variables.

Por último, cuando se calcula la *mediana*, también suelen estirmarse estadísticos que midan su representatividad en la distribución. Concretamente, la *desviación media de la mediana* y el *coeficiente de variabilidad de la mediana de Pearson* (obtenido del cociente entre la *desviación media de la mediana* y la *mediana)*. Su interpretación es análoga a los estadísticos anteriores.

Para un conocimiento más detallado, remito a la consulta de cualquier manual de estadística.

• *Medidas de la forma de la distribución*

Este término comprende la disposición espacial de los *valores* en una distribución: su *asimetría* y *curtosis*.

1) La *asimetría* constituye un indicador de la *agrupación* de las *frecuencias* en la *curva* de una distribución. Cuando su valor es cero, expresa que la *curva* es *simétrica;* es decir, que coinciden los valores de la *media,* la *moda* y la *mediana*. Si es mayor de cero, significa que la distribución es *asimétrica a la derecha (o de sesgo positivo)*. Los casos se agrupan a la izquierda de la *curva*. Esto significa la mayor presencia en la distribución de *valores* inferiores a la *media*.

   En cambio, si el valor del coeficiente de sesgo es inferior a cero, denota que la agrupación de los *valores* se produce a la derecha de la *curva*. Por lo que habría una mayor representación de los *valores* superiores a la *media*.

   En general, un valor superior a 0.8 (positivo o negativo) indica que la *asimetría* de la variable es importante.

2) La *curtosis* hace referencia a la mayor o menor *concentración* de *valores* en torno a la *media* de la distribución. Si existe una elevada concentración, la distribución será *leptocúrtica*. Esto significa que la *media* es muy representativa, al haber muy poca dispersión respecto a ella.

   La distribución será *platicúrtica,* en el caso contrario: cuando existe una elevada *dispersión* de *valores* respecto a la *media*. Lo que expresa su escasa representatividad.

Con estos estadísticos lo que se pretende es comprobar si existen valores muy extremos en una distribución. En caso de existir, puede decidirse su transformación, con la finalidad de alcanzar una distribución que se aproxime a la *normal (simétrica* y *mesocúrtica)*.

Tras los primeros *análisis univariados,* el investigador adquiere una descripción inicial de cada una de las *variables* que componen la investigación. Este conocimiento le

puede llevar a introducir *modificaciones en la configuración de las variables,* de forma que se faciliten los análisis posteriores. Estas *modificaciones* consisten, esencialmente, en agrupaciones de *variables* o de sus *atributos*.

Cuando en una distribución se observa que una o varias *categorías* presentan una *frecuencia* muy baja, habría que proceder a su *agrupación* con otras *categorías* de similar significado (si se pretende, con ellas, efectuar *análisis bivariables* o *multivariables).* Si la variable es *nominal,* la agrupación se produce con *categorías* que presentan un nexo común. Pero, si la variable es *ordinal,* de *intervalo,* o de *razón,* han de agruparse sólo los *valores* que se hallen más próximos en la escala (por *ejemplo,* las categorías "extrema derecha" con "derecha").

A la nueva *categoría* (formada de la agrupación de dos o más categorías originales) habría que darle una nueva denominación, que resuma la variedad de *atributos* que comprende. De este modo las posibilidades de análisis, de variables con *atributos* de escasa representación en la *muestra,* se amplían; aunque la agrupación lleve consigo pérdida de información.

Igualmente, puede *agruparse variables* similares, con la finalidad de componer una medida única que sintetice la información contenida en *variables* análogas.

9.3. El análisis bivariable

Después de la realización de los *análisis exploratorios,* procede la realización de *análisis bivariables,* tanto con fines *descriptivos* (describir al conjunto de la población observada), como *explicativos* (analizar posibles relaciones causales entre dos variables: la independiente y la dependiente).

9.3.1. Las tablas de contingencia

En la investigación social, la práctica habitual es la confección de *tablas de contingencia,* formadas del cruce, al menos, de dos *variables*. Estas *tablas* generalmente se obtienen mediante los comandos CROSSTABS y TABLES, en la mayoría de los programas estadísticos.

En estos comandos se especifican las *variables* a cruzar. Si puede establecerse una *relación causal* entre ellas, el orden convencional de exponer las *variables* es: *dependiente* "by" *independiente*. De esta forma, la *variable dependiente* figurará en las *filas,* mientras que la *independiente* en las *columnas*.

Si se desean introducir *variables de control* (para la eliminación de *explicaciones alternativas),* estas terceras y, a veces, cuartas variables, se añaden a las anteriores, siendo igualmente precedidas por la preposición "by" (por *ejemplo,* CROSSTABS = EDAD BY HERMANOS; CROSSTABS = MEDIDA BY DELITO BY SEXO). No obstante, en el programa informático que se emplee, se especifican las instrucciones propias para la ejecución del comando.

Además de las *variables,* ha de delimitarse la información que se precise: *frecuencias absolutas, porcentajes* (horizontales, verticales, totales), y *estadísticos* que midan el grado y la significatividad de la relación entre las *variables.*

El Cuadro 9.4 muestra una *tabla de contingencia simple* obtenida mediante el comando CROSSTABS de SPSS. En él se solicitaron las *frecuencias absolutas ("count"),* y las *relativas:* porcentajes horizontales *("row")* y verticales *("column"),* para cada *casilla* de la *tabla.* A ello se sumaron los *estadísticos de contingencia.*

CUADRO 9.4. Una tabla de contingencia simple mediante SPSS.

Crosstabulation: Delito By medida

Medida / Delito	Count / Row Pct / Col Pc	Amonestación 1	Libertad vigilada 2	Internamiento 3	Row total
Hurto	1	40 51.3 19.3	26 33.3 11.7	12 15.4 12.2	78 14.8
Intimidación con armas	2	12 19.7 5.8	36 59.0 16.1	13 21.3 13.3	61 11.6
Robo sin intimidación	3	116 45.8 56.0	93 36.8 41.7	44 17.4 44.9	253 47.9
Insumisión paterna	4	10 13.7 4.8	43 58.9 19.3	20 27.4 20.4	73 13.8
Otras infracciones	5	29 46.0 14.0	25 39.7 11.2	9 14.3 9.2	63 11.9
	Column total	207 39.2	223 42.2	98 18.6	528

Chi-Square	D.F.	Significance	Min E.F.	Cells with E.F.< 5
41.32788	8	.0000	11.322	None

Statistics	Symmetric	With DELITO Dependent	With MEDIDA Dependent
Lambda	.07069	.00000	.13443
Uncertainty Coefficient	.03488	.03033	.04102
Somers'D	.04951	.05219	.04710
Eta		.07227	.17561

Statistics	Value	Significance
Cramer's V	.19783	
Contingency Coefficient	.26943	
Kendall's Tau B	.04958	.0952
Kendall's Tau C	.04959	.0952
Pearson's R	.04051	.1764
Gamma	.07346	

Number of Missing Observation = 0

Cada *casilla* es el resultado del cruce de una *fila* con una *columna* (es decir, del cruce de un *atributo* de una *variable* con el *atributo* de la otra *variable).*

La lectura de las *tablas* con frecuencia se limita a comentarios porcentuales. Se contrastan los *porcentajes* de cada *casilla* para comprobar la existencia de variaciones entre los distintos *atributos* de las *variables.* A tal fin se calculan los *porcentajes* a partir de los *marginales* de la *tabla.*

Si se toma como base el total de *filas,* el *porcentaje* será *horizontal,* y las comparaciones porcentuales (entre los subgrupos) se efectuarán verticalmente. Por el contrario, cuando la base la constituye el total de *columnas*, se procede a la inversa: el *porcentaje* será *vertical,* y las comparaciones de porcentajes en sentido horizontal.

El investigador deberá escoger entre uno u otro tipo de *porcentaje* (horizontal o vertical), en conformidad con los *objetivos* del estudio y las *hipótesis* que compruebe. Si bien, existen mayores restricciones en *estudios explicativos* que en los *descriptivos.*

Si la finalidad de la investigación es la búsqueda de *relaciones causales,* los *porcentajes* se estiman sólo en el sentido de la variable *independiente.* Esta variable suele situarse en las *columnas,* y la *dependiente* en las *filas;* salvo que el elevado número de *atributos* de la variable *independiente* desaconseje su ubicación en las *columnas.* Esta disposición de las variables responde a la mayor facilidad de lectura (en la cultura occidental) en sentido horizontal frente al vertical. Los *porcentajes* serían, por tanto, *verticales* y las comparaciones horizontales.

Las diferencias porcentuales deben superar un determinado valor (al menos superior al 5%) para que puedan considerarse importantes. Depende del *error muestral* que derive de los tamaños de las *bases* sobre las que se calculan los *porcentajes.* Si estos *tamaños muestrales* son bajos, la diferencia porcentual ha de ser superior, si de ella quiere deducirse la existencia de *asociación* entre las *variables.*

En la exposición de la *tabla* (en el informe de la investigación) ha de indicarse, explícitamente, la dirección en la que se han calculado los *porcentajes.* Como sólo se aportan datos porcentuales, se recomienda poner entre paréntesis las *bases* de los *porcentajes.* Ello ayudará a la interpretación de las diferencias que en ellos se observen.

También se aconseja encabezar la *tabla* con un *título* que describa, sucintamente, el contenido de la *tabla.* En el *título* han de especificarse las *variables* comprendidas en la *tabla* y su relación.

A modo de ejemplo, la *tabla de contingencia* expuesta en el Cuadro 9.4 puede transformarse, en el *Informe,* como se expone en el Cuadro 9.5. Si bien, ha de matizarse que ésta constituye una de las posibles alternativas.

Pese a la existencia de convencionalismos en el formato de las tablas, el investigador es libre de diseñar el formato que más se ajuste a su estilo particular, y al contenido de la *tabla.*

La *tabla* expuesta en el Cuadro 9.5 constituye una *tabla simple* porque en ella figuran sólo dos *variables.* Pero también suelen componerse *tablas complejas,* a partir de

CUADRO 9.5. Menores clasificados por medida del tribunal, según el tipo de delito (porcentaje vertical.

Medida	*Tipo de delito*					
	Hurto	*Robo con intimidación*	*Robo sin intimidación*	*Insumisión paterna*	*Otros delitos*	*Total*
Amonestación	51	20	46	14	46	39
Libertad vigilada	33	59	37	59	40	42
Internamiento	16	21	17	27	14	19
Total	100 (78)	100 (61)	100 (253)	100 (73)	100 (63)	100 (528)

la conjunción de varias *variables* (independientes y/o dependientes). En estos casos, los análisis se centran en las variaciones en los *valores* extremos de las *variables*.

La *lectura porcentual,* aunque ilustrativa, resulta, no obstante, insuficiente. Precisa del complemento de *estadísticos* que gradúen la *asociación* entre las *variables* y su *significatividad*. Este complemento adquiere un mayor protagonismo cuando de las *tablas* quiera deducirse una *relación causal*.

Dos *variables* se hallan relacionadas si sus *atributos* varían conjuntamente. Para la graduación de esta relación, se acude a alguno de los *estadísticos de contingencia,* dependiendo del nivel de medición de la variable. Aquí sólo se enumeran. Consúltese en un manual de estadística su formulación.

a) Si la variable es *nominal,* los estadísticos que miden el grado de *asociación* entre dos *variables* son: Phi cuadrado, "C" de Pearson, "V" de Cramer, "Q" de Yule, Lambda, Tau-Y de Goodman y Kruskal...; además del coeficiente "d" o de diferencia de proporciones.
b) *Variables ordinales:* Rho de Spearman, Tau-A, Tau-B y Tau-C de Kendall, Gamma de Goodman y Kruskal, "D" de Sommer, entre otros.
c) *Variables de intervalo:* a los estadísticos anteriores se suma el coeficiente de correlación producto-momento de Pearson.

Cada uno de estos estadísticos indican la fuerza y la dirección de la *asociación* entre dos variables. Su *signo* expresa la dirección de la *correlación* (positiva o negativa); mientras que el *valor* numérico (que oscila entre 0 y 1), la magnitud de la relación. De este modo:

a) El valor ".00" denota *inexistencia* de *asociación*.
b) "–1.00", *correlación perfecta negativa* (conforme aumenta el valor de la variable independiente, disminuye el valor correspondiente a la dependiente).

c) "1.00", *correlación perfecta positiva* (al incremento de la variable independiente le sigue el aumento, también, de la dependiente).

Una vez medida la relación entre las variables, se comprueba su *significatividad.* Los datos analizados siempre pertenecen a una *muestra,* de las múltiples posibles, que pueden extraerse de una misma *población.* Razón por la cual, se exige la comprobación de la *significatividad* de los estadísticos y sus posibilidades de *inferencia* a la *población.*

En las *tablas de contingencia,* se aplica el *test de la Chi-Cuadrado* (X^2). Este estadístico se obtiene de la comparación entre las *frecuencias observadas* (en la *muestra)* y aquellas que cabría esperar en caso de inexistencia de relación entre las variables. Su valor se compara con el *teórico* (aquel que figura en una tabla de la X^2), para unos grados de libertad determinados y un nivel de significación escogido por el investigador. El *nivel de significación* habitual es .05. Este supone una posibilidad de error en la estimación del 5%. Los *grados de libertad,* en cambio, vienen marcados por las dimensiones de la *tabla:* número de *filas* (i) y de *columnas* (j). Concretamente, g. l. = (i–1)(j–1).

Cuando el valor de la X^2 empírico (el obtenido en la *muestra)* supera al teórico (el marcado en una tabla de la X^2), se deduce la *significatividad* estadística de la relación observada entre las *variables.* En caso contrario (X^2 empírico < X^2 teórico), se desestima la relación bivariable, por su no *significatividad.* La relación se consideraría, entonces, meramente casual, debida a *errores muestrales.*

En la salida de ordenador (véase Cuadro 9.4, por ejemplo) el valor de la X^2 aparece acompañado con su *significatividad ("significance").* El valor que figura bajo este rótulo ha de ser inferior a .05 para que el valor de la X^2 sea significativo a un nivel de .05.

En cualquier manual de estadística en las ciencias sociales (como el de Blalock, 1978; García Ferrando, 1985; o Siegel, 1985) puede encontrarse una exposición detallada de estos y demás *estadísticos de contingencia.*

9.3.2. Otros análisis bivariables

Aparte de las *tablas de contingencia,* existen otras *técnicas de análisis bivariables,* como el de *regresión* y *varianza simple.* Ambas técnicas analíticas miden la relación de *dependencia* entre dos variables, si bien imponen mayores restricciones que las *tablas de contingencia.* Su cumplimiento exige que la *variable dependiente* sea *métrica* o cuantitativa. Ello determina su menor aplicabilidad en la investigación social, en la que predominan las *variables cualitativas (no métricas).*

- *Varianza simple*

Esta técnica analítica es muy aplicada en los *diseños experimentales,* en la comprobación de los efectos de los tratamientos experimentales. A ello contribuye la es-

pecificidad del análisis. Su finalidad es comprobar la existencia de diferencias grupales respecto a una única *variable dependiente (métrica)*. Para ello se manipula una *variable independiente,* en función de cuyos valores se forman distintos grupos de tratamiento. Constituidos los grupos, se comprueba la media de cada uno de ellos respecto a la *variable dependiente*. Si se observan diferencias entre las *medias grupales,* se procede después a la comparación de las *varianzas grupales,* y a la medición de su *significatividad*.

Interesa que la *varianza* entre los grupos supere a la *varianza intragrupal*. Ello expresaría una mayor heterogeneidad entre los grupos, frente a una escasa variabilidad dentro de ellos. Por lo que podría afirmarse la existencia de diferencias entre los grupos.

La *significatividad* de las diferencias grupales se comprueba mediante los estadísticos "t" (si únicamente se han formado dos grupos de tratamiento), y "F" (si son más de dos los grupos creados).

En ésta, como en cualquier prueba de *significatividad*, se comparan los *valores empíricos* ("t" y "F") con los *teóricos* (mostrados en las tablas de la *"t" de Student* y de la *"F" de Fisher*, correspondientes). El proceso es similar al descrito en el *test de la Chi-Cuadrado*. Se fija el *nivel de significatividad,* en función de la precisión que el investigador desee para su estimación (.05, generalmente); y los *grados de libertad* (ahora determinados por el *tamaño muestral* y el número de *variables independientes* consideradas). Siempre que el *valor empírico* supere al *teórico,* las diferencias grupales observadas en la *muestra* adquirirán *significatividad estadística*. Podrán, por tanto, hacerse extensibles al *universo* del que se extrajo la *muestra* (en los niveles de probabilidad fijados).

• *Regresión simple*

Constituye otra técnica de *dependencia* en la que se analiza la relación entre una única variable *independiente (métrica o no métrica)* y una *dependiente (métrica)*. Pero, a diferencia de la técnica analítica anterior, la finalidad del análisis es la *predicción* del valor de la variable *dependiente* a partir del conocimiento de la *independiente*. Se cuantifica la relación existente entre ambas variables; y, se establece el grado de confianza o *significatividad* de la estimación efectuada.

La *correlación* entre las dos variables *(dependiente* e *independiente)* se mide mediante el *coeficiente R de Pearson*. Éste expresa el grado de *covariación* entre las variables, según se aproxime a "0" (inexistencia de asociación) o a "1" (asociación perfecta). También informa de la dirección de la asociación: creciente (si el signo es positivo) o decreciente (si es negativo).

En el *análisis de regresión,* la idea que subyace es la consecución de una *recta de regresión* que presente el mejor "ajuste" de los casos respecto a las variables analizadas. Esta *recta* tiene su expresión matemática en la siguiente *ecuación de regresión:*

$$y = a + bx + e$$

donde: "y" denota el valor de la variable *dependiente.*
"a" es el *intercepto* o punto de la *recta* que corta al eje de las Y.
"b" es la *pendiente* de la *recta* (también referido como el *coeficiente de regresión).* Su *valor* expresa la cantidad de variación de la variable *dependiente* por cada unidad de variación de la *independiente.* Su *signo* denota si se produce aumento (pendiente creciente; signo positivo) o disminución (pendiente decreciente; signo negativo).
"e" representa el *error de la estimación:* la inadecuación de la *ecuación de regresión* en la predicción del valor de la variable *dependiente.*

Esta *ecuación* permite la predicción del valor de la variable *dependiente* a partir de valores conocidos de la *independiente.* Los *coeficientes* se obtienen, generalmente, siguiendo el *criterio de mínimos cuadrados* (hacer mínima la distancia que separa los *puntos* –obtenidos de la confluencia de ambas variables en cada uno de los casos– y la *recta de regresión).*

La *significatividad* de los *coeficientes* se comprueba mediante el estadístico "*t*", con n-1 *grados de libertad* (siendo *"n"* el número de observaciones). En cambio, la *significatividad* de la *correlación* se comprueba mediante el estadístico "*F*". Como en cualquier prueba de *significatividad,* los valores de "*t*" y de "*F*" *empíricos* han de superar los *teóricos* (determinados en las tablas respectivas) para que el modelo de *regresión* sea significativo estadísticamente.

9.4. El análisis multivariable

Los análisis *univariables* y *bivariables* con frecuencia se muestran insuficientes para cubrir los objetivos de la investigación. El proporcionar una visión conjunta e integrada, que describa y/o explique la realidad que se analiza, demanda la realización de *análisis multivariables* (de más de dos variables al mismo tiempo). De otra forma no podrían medirse las influencias e interrelaciones existentes entre grupos de variables. Como hace tiempo reconociera García Ferrando (1979: 198):

> "Las distribuciones bivariables en sociología aparecen demasiado simplistas para lograr adecuadas explicaciones científicas."

La peculiaridad del *análisis multivariable* reside en operar con un número elevado de variables, y de manera simultánea, basándose en el cálculo matricial. Kendall (1975) lo define como el conjunto de técnicas estadísticas que permite el análisis simultáneo de más de dos variables en una muestra de observaciones. A esta definición, Dillon y Goldstein (1984) añaden la posibilidad de análisis sincrónicos de mediciones en más de una *muestra.*

Para el *análisis multivariable* existe un amplio abanico de *técnicas*. En conformidad con un extenso grupo de autores (véase Kendall, 1975; Dillon y Goldstein, 1984; o Hair *et al.* 1992, por ejemplo), estas técnicas pueden agruparse en dos grandes categorías *(técnicas de dependencia y técnicas de interdependencia),* en función de si se diferencia. o no, entre variables *dependientes* e *independientes*.

En la elección de la *técnica* concreta a aplicar intervienen, básicamente, el objetivo de la investigación, y las características de las variables que se analicen (su número y nivel de medición).

A continuación se ofrece una breve descripción de cada una de las *técnicas multivariables* comúnmente referidas. Para un mayor conocimiento de cada una de ellas remito a la bibliografía especializada. Lo que sigue es una mera visión panorámica de las *técnicas multivariables*.

9.4.1. Técnicas multivariables de dependencia

Un conjunto de técnicas analíticas unidas por un mismo propósito: medir la existencia de *relaciones causales* entre un conjunto de variables, el grado y *significatividad* de la misma. Sin embargo, difieren en el número de variables *dependientes* que incluyen, y en el nivel de *medición* exigido *(métrico o no métrico).* En la Figura 9.1 se esquematizan las principales alternativas, comúnmente señaladas, en el análisis de la *dependencia.*

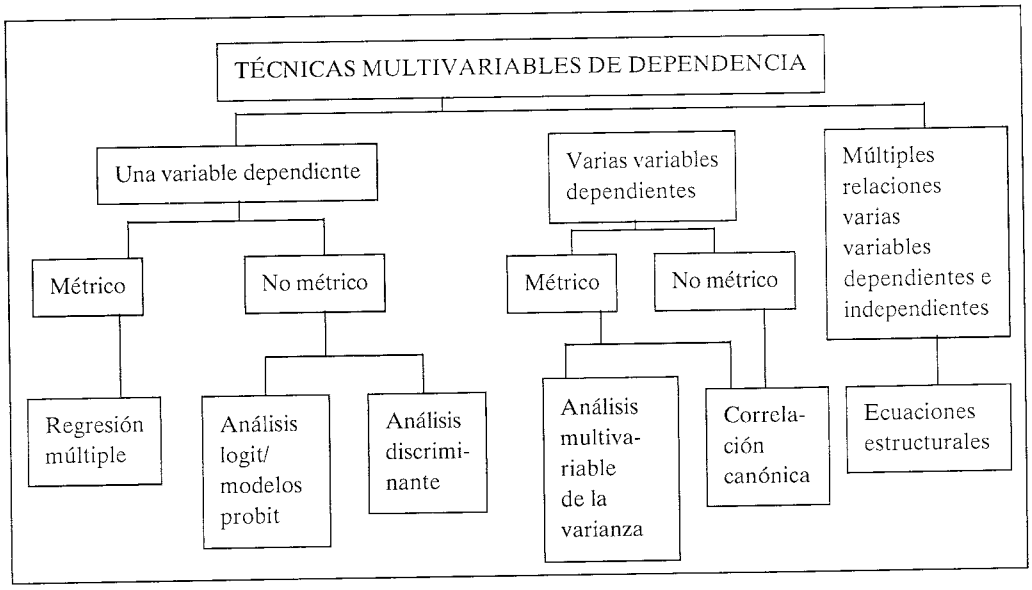


Figura 9.1. Técnicas de análisis multivariable de dependencia.

• *Regresión múltiple*

Una técnica de *dependencia* muy versátil, y utilizada, cuando se busca la *predicción* del valor de una variable *dependiente* (o *criterio),* a partir de valores conocidos en una serie de variables *independientes* (o *predictoras).* Para su realización se exige el cumplimiento de una serie de *supuestos* básicos:

a) *Linealidad:* la interdependencia entre las variables ha de responder a un modelo lineal. Esto quiere decir, que a cada cambio en una unidad en una variable corresponda un cambio, también en una unidad, en la otra variable.

b) *Normalidad:* la correspondencia de la distribución de los datos (para la variable *dependiente* y las *independientes)* con la *curva normal.* Esto permite la utilización de los estadísticos *"F"* y *"t"* para la comprobación de la *significatividad.*

c) *Homocedasticidad* o igualdad de las *varianzas* de los términos de error en la serie de variables *independientes.*

Para que pueda medirse la relación de las variables *independientes* con la *dependiente* se precisa que la *varianza* de los valores de la variable *dependiente* sea igual en cada valor de las variables *predictoras.*

d) *Aditividad:* los efectos de las variables *independientes* han de poderse sumar entre sí, para poder predecir la variable *dependiente.*

e) Ausencia de *colinealidad* (de correlación) entre las variables *independientes,* con objeto de que puedan medirse sus efectos concretos en la variable *dependiente.*

f) Ha de haber un número elevado de observaciones. Al menos, debería haber 20 veces más casos que variables *independientes.* Máxime si el modelo de regresión se obtiene "paso a paso": incorporando una a una las variables *independientes* hasta que no exista ninguna más con poder predictivo significativo.

Cuando no se obtiene esta proporción, habría que optar por alguno de los siguientes remedios: eliminar alguna variable *independiente,* o agrupar varias variables creando una nueva variable (que sea una combinación de dos o más variables *independientes).*

Como en *regresión simple,* en *regresión múltiple* se obtiene una *ecuación de regresión,* con tantos *coeficientes* como variables *independientes.* A partir de ella, podrá predecirse el valor medio de la variable *dependiente.* Asimismo, se mide el grado de *correlación* existente entre las variables (mediante el estadístico *R múltiple)* y su *significatividad* (gracias al estadístico *"F").*

El valor del R^2 *múltiple* (el *coeficiente de determinación)* expresará el porcentaje de *varianza* de la variable *dependiente* que es explicado por las *independientes.* Cuanto más elevado sea su valor, mejor para la predicción de la variable *dependiente.*

• *Análisis logit*

Los análisis *logit* y *probit* se definen como modelos analíticos alternativos (al modelo de *probabilidad lineal)* para variables *dependientes cualitativas.* Ambos modelos *(logit* y *probit)* se asemejan bastante. Si bien, el *análisis logit* ha alcanzado un mayor desarrollo, en parte debido a su mayor practicabilidad. De ahí que se le conceda un mayor protagonismo en esta breve reseña analítica.

El *análisis logit* constituye un modelo de respuesta cualitativa, que mide la relación entre una serie de variables *independientes (métricas* o *no métricas)* y una única variable *dependiente categórica (no métrica).* Para ello se basa en el análisis de la *razón de probabilidad* de una variable *dependiente,* en función de las *independientes.* De lo que se trata es de graduar la *probabilidad* de que determinadas variables *independientes* provoquen la ocurrencia de un evento concreto (la variable *dependiente).*

Del análisis se obtiene una *ecuación* similar a la de *regresión.* Los *coeficientes* (calculados siguiendo el criterio de *máxima verosimilitud,* o el de *mínimos cuadrados)* se interpretan como en *regresión.* La *significatividad* de cada uno de ellos también se comprueba mediante el estadístico *"t".*

La peculiaridad del análisis reside, no obstante, en el cálculo de los *incrementos de probabilidad.* Éstos se calculan para cada variable con *coeficiente* significativo. Se considera el valor del *coeficiente* (ß) y el *valor medio* de la variable en la *muestra* (X_j), siguiendo la expresión de McFadden (1974):

$$Prob(Y=1)=F(X'_J\beta)=\frac{e^{X'_j\beta}}{1+e^{X'_j\beta}}$$

La *significatividad* del modelo global se comprueba mediante el estadístico X^2. Este estadístico se complementa con el *porcentaje de aciertos* (el porcentaje de observaciones que logran ser correctamente estimadas mediante los *coeficientes logit),* como un indicador del éxito del modelo en la predicción de la variable *dependiente.*

• *Análisis discriminante*

Una técnica de clasificación y de asignación de individuos a grupos, a partir del conocimiento previo de sus características.

Su objetivo básico es la estimación de la relación existente entre una serie de variables *independientes* y una única variable *dependiente no métrica* (o *categórica).*

El número de *categorías* que incluya la variable *dependiente* determina los grupos formados a partir de la *muestra.* Estos *grupos* deben, previamente, haberse configurado mediante la aplicación de una *técnica multivariable de interdependencia* (como el análisis de *conglomerados* o el *factorial).*

Con el *análisis discriminante* lo que se pretende es la comprobación de si los *grupos* creados por otra técnica analítica son adecuadamente caracterizados por las variables que les definen (de acuerdo con los análisis previos).

Asimismo, se trata de conocer cuál es la combinación de variables *(funciones discriminantes)* que hace máxima la diferenciación entre los grupos. El conocimiento de estas variables ayudará a la predicción de la probabilidad de pertenencia de un individuo concreto a uno de los grupos diferenciados. Ello dependerá de los valores que presente en las variables *independientes* analizadas.

La ejecución del *análisis discriminante* exige, igualmente, el cumplimiento de unos *supuestos* claves, como son:

a) La *normalidad* de las variables *independientes*. Se recomienda que éstas sean *métricas*. En caso contrario, habría que introducir modificaciones en el análisis.
b) *Matrices de covarianzas* iguales en cada grupo.
c) Inexistencia de *colinealidad* entre las variables *independientes,* para evitar que su información sea redundante en la explicación de la variable *dependiente.*
d) Ha de haber, al menos, 20 casos por cada variable *independiente* introducida en el análisis. Esto contribuye a la *significatividad* estadística del modelo obtenido.

La distancia entre los grupos se mide mediante el estadístico D^2 *de Mahalanobis*. A éste se añaden los estadísticos *lambda de Wilks, F* y X^2, en la comprobación de la *significatividad* de las variables *independientes* en la diferenciación entre los grupos.

La contribución de estas variables se cuantifica mediante los *coeficientes de función discriminante estandarizados* (similares a los *coeficientes beta* en *regresión)* y los *factores de carga discriminantes* (los *"discriminant loadings").* Sobre todo, por estos últimos, debido a su mayor precisión.

Como en el *análisis logit,* en el *discriminante* también se calcula la *razón de aciertos*, como medida del éxito del modelo en la predicción de la variable *dependiente* (la clasificación de los grupos).

• *Análisis multivariable de la varianza*

También conocido como MANOVA. Representa una extensión del análisis univariado de la varianza (ANOVA), mediante el cual se exploran, simultáneamente, las posibles relaciones existentes entre: varias variables *independientes no métricas* (normalmente referidas como *tratamientos)* y dos o más variables *dependientes métricas*. Lo que permite la medición de las *correlaciones* entre las variables *dependientes* y entre las *independientes*.

El uso de este análisis se adecúa, igualmente, a los *diseños experimentales,* en la comprobación de los efectos de distintos *tratamientos*. El fin que se pretende es la determinación de la existencia de diferencias, en los *valores medios* de las variables *dependientes,* en cada *grupo de tratamiento*. Ello exige que:

a) Las variables *dependientes* se hallen distribuidas *normalmente.*
b) Se parta de *matrices de varianza-covarianza* iguales en todos los *grupos.* En caso contrario, no podría medirse el efecto específico del *tratamiento* dado (debido a que se partiría de grupos inicialmente diferentes).
c) El *tamaño muestral* ha de superar al necesario para la *varianza simple,* si se pretende que el modelo adquiera *significatividad* estadística.

La *significatividad* de las diferencias multivariables entre los *grupos* se comprueba mediante cuatro criterios primordiales: la *raíz máxima de Roy,* la *Lambda de Wilks,* la *traza de Hotelling* y la de *Pillai.* Si estos tests multivariados resultan significativos, se procede a la comprobación de las diferencias grupales en cada una de las variables *dependientes.* Para ello se acude al estadístico "*F*".

• *Correlación canónica*

Una *técnica de dependencia* que permite la comprobación de la existencia de interrelación entre una serie de variables *dependientes* y otra serie de variables *independientes.* Ambas pueden ser tanto *métricas* como *no métricas.*

La finalidad principal del análisis es la obtención de *combinaciones lineales* de cada serie de variables *(dependientes* e *independientes).* Estas combinaciones han de maximizar las *correlaciones* entre las variables.

El grado de relación entre la serie de *variables canónicas* se mide mediante el *coeficiente de correlación canónica.* Éste, elevado al cuadrado (R^2), representa el porcentaje de *varianza* de una combinación de variables *dependientes canónicas* que es explicada por una combinación de las variables *independientes.* Su *significatividad* se comprueba, de nuevo, mediante el estadístico "*F*".

Las contribuciones de cada variable (*dependiente* e *independiente*) a la combinación o serie de *variables canónicas* respectiva se comprueba, al igual que en el *análisis factorial* y *discriminante,* mediante los *factores de carga ("canonical loadings").* Estos han de ser ≥.30 para que se consideren relevantes. La contribución de las variables también puede comprobarse mediante los *pesos canónicos ("canonical weights"),* si bien estos últimos presentan una mayor inestabilidad que los anteriores (los *factores de carga).*

• *Ecuaciones estructurales*

También conocidas como modelos LISREL, en referencia a uno de los programas estadísticos más populares para su resolución.

Esta última *técnica de dependencia* puede catalogarse como una extensión del *análisis factorial* y de *regresión múltiple,* por dos razones fundamentales:

a) En las relaciones de *dependencia* se representan variables *latentes* o constructos (no observadas), a partir de valores conocidos en las variables *manifiestas* (u observadas) –a semejanza con el *análisis factorial*–.
b) Las diversas *relaciones causales* tienen su expresión matemática en varias *ecuaciones de regresión* (ahora denominadas *ecuaciones lineales estructurales).*

Pero, a diferencia de las *técnicas de dependencia* anteriormente expuestas, en los *modelos de ecuaciones estructurales* se comprueba (simultáneamente) diversas *relaciones causales*. No sólo entre varias variables *independientes (métricas o no métricas)* y *dependientes (métricas),* sino entre ellas mismas, también (dependientes con dependientes e independientes con independientes). De ello resulta un *modelo causal* de mayor complejidad, aunque más ajustado a la pluridimensionalidad de las *relaciones causales*.

Las *relaciones causales* también se representan gráficamente, mediante un diagrama de "sendero" *("path").* En él se diferencian las variables *latentes* de las *observadas*, además de los *errores de medición* (de cada modalidad de variable), y el tipo de relación entre ellas (por ejemplo, una flecha bidireccional indica la existencia de correlación entre dos variables).

Para cada variable que recibe una flecha se formula una *ecuación*. En esa *ecuación,* dicha variable actúa como variable *dependiente*.

La realización de esta técnica analítica exige el cumplimiento de los mismos supuestos que en las anteriores (como la *independencia* de las observaciones; la *linealidad* de las relaciones; la *normalidad* multivariable; la selección *aleatoria* de las unidades muestrales; y un elevado *tamaño muestral).*

Los *parámetros* se obtienen mediante el método de *mínimos cuadrados generalizados,* o por el de *máxima verosimilitud*. Si bien, el primer método muestra imprecisión, conforme desciende el tamaño de la *muestra* y aumenta la complejidad del modelo. La *significatividad* de los *parámetros* se comprueba mediante la *"t",* además de considerarse el *error* de la estimación.

Como se diferencia entre variables *latentes* y *observadas,* se comprueba la *fiabilidad* (≥.70) y la *validez* (≥.50) de la *medición* de los *constructos teóricos (variables latentes).* A ello se suma el *ajuste* global del modelo alcanzado respecto al inicial.

Existen distintos *índices de ajuste*. Entre ellos destacan los siguientes: el estadístico X^2 (aunque ahora interesan los valores bajos, porque expresan un mejor *ajuste* entre el modelo obtenido y el propuesto inicialmente); el *índice GFI*, que oscila entre .00 y 1.00 (cuanto más se aproxime a 1, mejor); los *índices TLI* (≥.90), *NFI* (≥.90), *AGFI* (≥.90). y *AIC* (≥.70), entre otros.

9.4.2. Técnicas multivariables de interdependencia

A diferencia de las técnicas analíticas anteriores, las de *interdependencia* presentan un menor poder predictivo. Mediante ellas se analiza la existencia de *asociación* o

relación mutua entre varias variables, sin diferenciar entre *dependientes* e *independientes*.

La Figura 9.2 representa las principales técnicas de interdependencia. La elección entre una u otra responde a los mismos criterios fundamentales mencionados en las *técnicas de dependencia*. Concretamente, el objetivo de la investigación y el nivel de medición de las variables. Ahora la diferencia básica entre las técnicas se establece en función del nivel de *medición* mínimo exigido en las variables para su cumplimentación: *métrico* o *no métrico*.

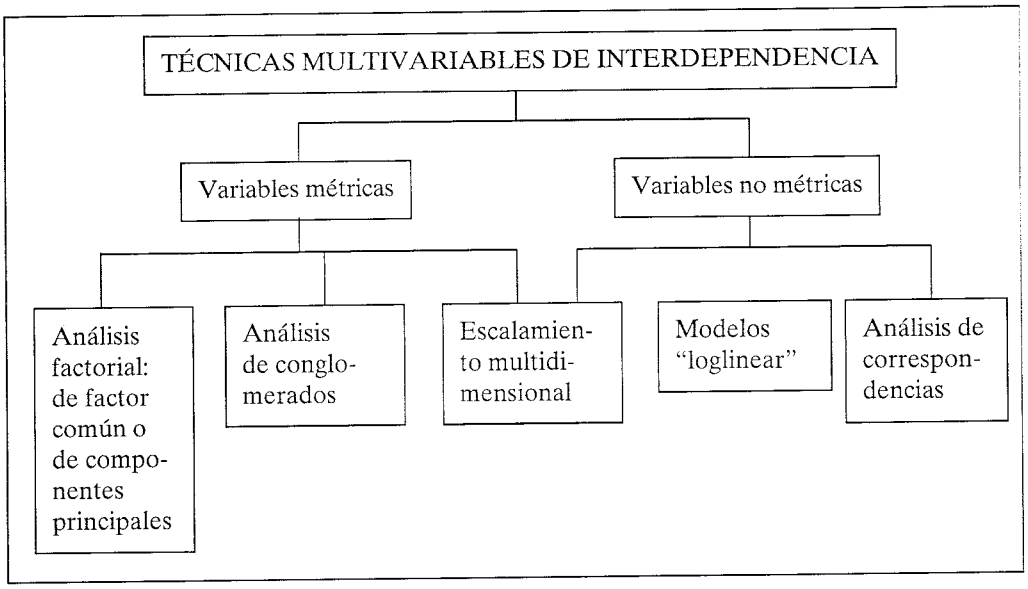


Figura 9.2. Técnicas de análisis multivariable de interdependencia.

• *Análisis factorial*

Una denominación genérica que engloba distintos procedimientos centrados en el análisis de la variación total que una variable comparte con otras variables. Su objetivo fundamental es resumir, la información contenida en un conjunto de variables interrelacionadas, en un número reducido de *dimensiones latentes* comunes (o *factores).*

El *análisis factorial* es *exploratorio,* cuando no se conoce, previamente, cuáles son los *"factores"*. Éstos se determinarán, precisamente, tras el *análisis factorial*. En cambio, el análisis será *confirmatorio,* si se parte de unos *"factores"* especificados *a priori,* mediante un análisis *factorial exploratorio* u otra técnica multivariable (como el análisis de *conglomerados).* Estos *factores* representan a un conjunto de variables *empíricas* u observadas. Entonces, se tratará de "corroborar" la adecuación de estas variables (empíricas) en la medición de las *dimensiones* de los *conceptos teóricos*.

Además de esta clasificación básica, se diferencia entre análisis factorial de *componentes principales* y de *factor común*. En el *análisis de componentes* se persigue la obtención de *combinaciones lineales* de variables que logren explicar la mayor proporción de *varianza* conjunta. Para ello se tiene en cuenta tanto la *varianza específica* de cada variable, como la *varianza conjunta* (o compartida con otras variables). Por el contrario, en el *análisis de factor común,* la identificación de los *factores latentes* responde sólo a la *varianza común* de una serie de variables.

En cualquiera de las modalidades de *análisis factorial,* el investigador deberá concretar el método para la extracción de los *factores: ortogonal* (quartimax, varimax o equimax) u *oblicuo*. En el *ortogonal,* los *factores* se extraen de manera que sean totalmente independientes unos de otros. La extracción *oblicua* permite, en cambio, la existencia de correlación entre los *factores*.

En la decisión de cuántos *factores* escoger han de compaginarse distintos criterios como, por ejemplo, el criterio de *raíz latente (autovalor* superior a 1), o el de *porcentaje acumulado de varianza* (al menos superior al 60%).

Las *correlaciones* de las variables *empíricas* con los *factores* se comprueban mediante los *factores de carga ("factor loadings").* Éstos han de ser ≥.30 para considerarse significativos.

Una vez definidos los *factores,* se les asigna un nombre o *etiqueta*, que refleje el contenido de las variables *empíricas* que representa.

Las investigaciones comentadas en los Capítulos 4 y 10 ilustran la aplicación de esta técnica analítica. Razón por la que se remite a su lectura.

• *Análisis de conglomerados*

Una variedad de análisis cuya finalidad principal es la clasificación de un grupo de individuos u objetos en un número reducido de grupos. Estos *grupos* han de ser mutuamente excluyentes; han de estar compuestos por individuos lo más similares posible entre sí y diferentes de los integrantes de otros grupos.

A diferencia del *análisis discriminante,* los *grupos* no se hallan predefinidos. Precisamente se aplica esta técnica analítica para la definición de los grupos. Posteriormente, éstos pueden ser corroborados mediante otras técnicas multivariables (como el análisis *discriminante* o el *factorial confirmatorio).*

El *grupo* se define en función del algoritmo de clasificación que se emplee en la agrupación de los sujetos. En general, se busca la agrupación que haga máxima la distancia entre las *medias grupales* y mínima la *desviación intragrupal* (de los individuos integrantes del grupo respecto a su *centroide* o *media grupal).*

Para la obtención de estos grupos, o *conglomerados,* puede elegirse entre distintos procedimientos, agrupados en dos amplias categorías: *jerárquicos* y *no jerárquicos.* Los primeros se dividen, a su vez, en *aglomerativos* y *disociativos* (o *divisorios).*

a) El procedimiento *jerárquico aglomerativo* comienza con tantos grupos como individuos en la *muestra*. Paulatinamente van reduciéndose el número de grupos

y aumentando, en contra, el número de individuos en cada grupo. El programa finaliza cuando un único *conglomerado* agrupa a todos los individuos.

b) El procedimiento *jerárquico disociativo* (o *divisorio)* procede a la inversa. Parte de un único grupo, que integra a todos los individuos. Poco a poco, este grupo va seccionándose en diferentes subgrupos hasta que, al final, existen tantos grupos como individuos.

En ambos procedimientos *jerárquicos* de formación grupal, la elección del número de *conglomerados* se realiza conforme a una variedad de criterios. Entre ellos destacan: el criterio de la *distancia mínima;* de la *distancia máxima;* el *promedio de las distancias;* el *método Ward;* o el de los *centroides.* Mientras que los primeros criterios consideran la distancia de los individuos, el último tiene en cuenta la distancia entre las *medias (centroides)* de las variables. En cualquiera de estos procedimientos *jerárquicos* los datos pueden visualizarse mediante un *dendograma:* una representación gráfica en forma de árbol.

En los procedimientos de agrupación *no jerárquica* se parte, a diferencia de los anteriores, de una especificación previa de los grupos que desean formarse. De lo que se trata es de encontrar representantes para cada uno de los grupos. Un individuo pertenecerá al grupo cuya distancia a su *centroide* sea menor.

Una vez que los grupos se han constituido (por cualquiera de los procedimientos referidos), se procede a la comparación de las *varianzas* de las variables en relación a los grupos *(intergrupo* e *intragrupo).* Ello permite comprobar si los grupos presentan diferencias en los valores de las variables consideradas en el análisis.

En función del valor del estadístico *F* (en cada una de las variables), se rechazan aquellas variables que no diferencian a los grupos. A partir del valor de la *media* de las variables se extrae, por último, los *rangos* que más separan a los integrantes de cada grupo.

• *Escalamiento multidimensional*

Una variedad analítica análoga al *análisis factorial.* Con él comparte un mismo objetivo: la obtención de un número reducido de *dimensiones,* que permitan caracterizar a determinados objetos o sujetos. Si bien difiere (entre otros aspectos) en el número de *dimensiones* a obtener. Mientras que el *análisis factorial* no impone restricciones al respecto (de hecho el análisis puede efectuarse con un número elevado de *dimensiones* o *factores),* el *escalamiento multidimensional* aconseja su reducción al menor número posible. Ello responde a condicionamientos impuestos para la representación gráfica de los resultados de la investigación.

En el *escalamiento multidimensional* los datos se representan como puntos en un espacio "multidimensional". La distancia habida entre ellos se considera una expresión gráfica de su semejanza o disimilaridad. Esta representación se hace más fac-

tible e interpretable cuando la información puede reducirse a dos o tres *dimensiones,* como máximo.

Para la concreción de estas *dimensiones* se acude a uno (o varios) *criterios de bondad de ajuste:* el "stress" de Kruskal, el "s-stress" de Young, la *correlación múltiple al cuadrado,* o el *diagrama de Shepard.*

Para que la configuración de las *dimensiones* presente un *ajuste* adecuado, los valores correspondientes a "stress" y "s-stress" han de ser bajos. En cambio, la *correlación múltiple al cuadrado* ha de ser elevada. Su valor se interpreta como proporción de *varianza* explicada por las distancias respectivas. Por su parte, el *diagrama de Shepard* ha de reflejar una tendencia ascendente (o creciente).

Dependiendo del nivel de medición de las variables, se diferencia entre *escalamiento multidimensional "métrico"* y *"no métrico".*

a) *Métrico,* cuando las variables son de *intervalo* o de *razón.* Lo que favorece su configuración en una escala *continua.*

A partir de una *matriz de correlaciones,* o de *distancias* entre objetos, se trata de situar a éstos en un espacio *multidimensional.* Para ello se transforma la similaridad o disimilaridad percibida en ellos, en distancias *(euclidianas).*

b) *No métrico,* si las variables son *cualitativas (nominales* u *ordinales).* En este caso, se parte de una *matriz de rangos.* De ésta se obtiene información de la similaridad de los objetos. La configuración final será aquella que mejor represente a los *rangos* de la *matriz* inicial.

Como en el *análisis factorial,* el *escalamiento multidimensional* puede ser *exploratorio* y/o *confirmatorio.* Depende de la finalidad del mismo.

• *Modelos "log-linear"*

También conocidos como modelos "lineales logarítmicos" o modelos "log-lineales". Su denominación deriva de la transformación logarítmica operada en los datos (las frecuencias observadas en las variables) para facilitar su aditividad: la posibilidad de sumar los distintos efectos de un conjunto de variables *no métricas* (a semejanza de las variables *métricas* en el *análisis de regresión).*

Las *tablas de contingencia* resultan muy útiles en el análisis de variables *cualitativas.* Pero, cuando se añaden terceras o cuartas variables, a modo de variables de *control,* la interpretación de la *tabla* se complica. En estos casos, sobre todo conforme aumenta el número de variables, se aconseja la aplicación de *modelos log-linear.* Éstos se muestran adecuados para el análisis de las interrelaciones entre una serie de variables *no métricas,* que conforman una *tabla de contingencia multidimensional.*

Las casillas de la *tabla* se traducen a componentes, denominados *parámetros lambda.* Estos miden el efecto de los valores de las variables que conforman cada casilla. Si

el valor Z correspondiente a cada *lambda* es ≥ 1.96, el *parámetro lambda* se considera relevante en la interpretación del modelo. De él se pretende obtener la probabilidad de que un individuo concreto comparta una combinación específica de atributos de un conjunto de variables.

Para la comprobación del *ajuste del modelo* se acude a los estadísticos X^2 *y la razón de verosimilitud.* Los valores de ambos estadísticos tienden a coincidir conforme aumenta el tamaño de la muestra.

• *Análisis de correspondencias*

Una de las *técnicas de interdependencia* de más reciente desarrollo, también adecuada al análisis de variables *cualitativas.*

Parte de la configuración de las variables en una *tabla de contingencia.* Su objetivo fundamental es la representación de las distancias de las *filas* y las *columnas,* que integran la *tabla,* en unos *ejes cartesianos.* Para ello se transforma el valor de la X^2 en una medida métrica de distancia.

Como en el *escalamiento multidimensional,* la proximidad de los puntos mide la similaridad existente entre ellos. En el *análisis de correspondencias,* la proximidad muestra, concretamente, la asociación entre las categorías de las variables.

Primero se procede (como en el *análisis factorial* y el *escalamiento multidimensional)* a la identificación del número adecuado de *dimensiones* que categoricen al objeto de estudio. A tal fin, se examina, igualmente, el porcentaje acumulado de *varianza* explicada.

Para cada *dimensión* se deriva, a continuación, unos *autovalores ("eigenvalues").* Éstos expresan la contribución relativa de cada *dimensión* en la explicación de las *varianzas* de las variables.

Una vez establecida la *dimensionalidad,* se identifica la *asociación* o relación existente entre las categorías de las variables mediante su *proximidad.* Ésta ha de comprobarse o en las *filas* o en las *columnas.* Depende de su ubicación. Un valor elevado de X^2 indica un fuerte grado de "correspondencia" entre los atributos de las variables.

De esta forma se obtiene la reducción dimensional de las proporciones de objetos en una serie de atributos. Al mismo tiempo, se extrae la representación de los objetos, relacionados con esos atributos.

9.5. Paquetes estadísticos disponibles

Sin la mediación del ordenador no sería viable ninguno de los análisis estadísticos reseñados. Especialmente, los análisis *multivariables.*

La gran revolución experimentada en los últimos años en el campo de la informática hace que la información que se dé quede obsoleta en un breve período de tiem-

po. Por esta razón, únicamente van a nombrarse algunos de los principales *paquetes estadísticos,* para conocimiento del lector.

La distinción entre los *paquetes estadísticos* responde a su especificidad. Se diferencia entre programas "genéricos" (que ejecutan la generalidad de los análisis estadísticos), y los "específicos" (especializados en técnicas analíticas concretas).

Entre los *paquetes estadísticos genéricos* destacan el SPSS, BMDP, SAS, ESP, y OSIRIS. En sus manuales respectivos se detallan tanto aspectos técnicos (de funcionamiento del programa), como analíticos (para la interpretación de los análisis estadísticos). Dicha interpretación será la misma, independientemente del programa utilizado.

Los *paquetes estadísticos específicos* son numerosos y variados. A modo de ejemplo se señalan los siguientes: LISREL y EQS (para el análisis de *ecuaciones estructurales),* LIMDEP (análisis *logit),* INDSCAL *(escalamiento multidimensional),* ECTA *(tablas de contingencia),* SPAD (análisis de *correspondencias),* o BROCOLI *(series temporales).*

A estos paquetes estadísticos se suman otros específicos para la realización de *gráficos* (como el SYSTAT, STATGRAPHICS, o el HARVARD GRAPHICS). También hay que mencionar las *bases de datos numéricas* (como DBASE o SIR-DB). Estas *bases* están diseñadas para la exportación automática de datos en la mayoría de los paquetes estadísticos (en especial, los *genéricos).*

Finalmente, cabe mencionar que en los últimos años comienzan a comercializarse *sistemas expertos en análisis de datos* (como el GLIM). Éstos funcionan a modo de consultor experto en *técnicas cuantitativas de análisis.*

Lecturas complementarias

Bisquerra, R. (1989): *Introducción conceptual al análisis multivariante.* Barcelona, PPU.

Garrido Luque, A.; J. L. Álvaro Estramiana (1995): *Técnicas de análisis estadístico en ciencias sociales.* Madrid, Universidad Complutense.

Hair, J. *et al.* (1992): *Multivariate data analysis.* New York, McMillan.

Sánchez Carrión, J. J. (comp.) (1984): *Introducción a las técnicas de análisis multivariable aplicadas a las ciencias sociales.* Madrid, CIS.

Sánchez Carrión, J. J. (1989): *Análisis de tablas de contingencia: el uso de los porcentajes en las ciencias sociales.* Madrid, CIS, *Monografía* n.º 105.

Sánchez Carrión, J. J. (1995): *Manual de análisis de datos.* Madrid, Alianza Universidad.

Sánchez Carrión, J. J. y Torcal, M. (1992): *Utilidades del SPSS/PC+. Presentación de informes. grabación de datos y creación de gráficos y mapas.* Madrid, Alianza.

Spiegel, M. (1991): *Estadística. Teoría y problemas resueltos.* Madrid, McGraw Hill.

Ejercicios Propuestos

1. Interprete los datos que figuran en los Cuadros 9.3. y 9.4.
2. Describa una investigación en la que se haya aplicado alguna técnica analítica multivariable. Especifique los análisis efectuados y su adecuación respecto a los objetivos de la investigación.
3. Trace el plan de análisis que seguiría para obtener el perfil del alumnado de la facultad, añadiendo las razones de su elección.
4. Si la investigación tuviera como objetivo comprobar qué variables influyen en la probabilidad de aprobar una asignatura, ¿qué análisis efectuaría? Justifique la respuesta.
5. A partir de los datos siguientes, confecciones una tabla e interprétela.

Edad	*Sexo*	*Ideología política*	*Frecuencia absoluta*
De 18 a 25 años	Varón	Izquierda	130
		Centro	220
		Derecha	275
	Mujer	Izquierda	115
		Centro	105
		Derecha	310
De 26 a 40 años	Varón	Izquierda	175
		Centro	270
		Derecha	210
	Mujer	Izquierda	196
		Centro	307
		Derecha	150
De 41 a 60 años	Varón	Izquierda	320
		Centro	178
		Derecha	97
	Mujer	Izquierda	205
		Centro	240
		Derecha	170

10

EL ANÁLISIS DE CONTENIDO CUANTITATIVO

En el análisis de *documentos* (y, en general, de cualquiera de las modalidades de comunicación verbal, e incluso visual) adquiere un destacado protagonismo el *análisis de contenido*. Aunque algunos autores (Krippendorf, 1980, 1990; y algunos de sus seguidores: López Aranguren (1989), O'Connell y Layder (1994), entre otros) lo cataloguen como "técnica de investigación", considero que las características del *análisis de contenido* corresponden más bien a una "técnica de análisis", de utilidad en cualquiera de las *estrategias de investigación* anteriormente referidas. Desde el *uso de fuentes documentales* y el *estudio de casos,* hasta la propia *encuesta* (en la *codificación* de *preguntas abiertas).*

La posibilidad de utilización de esta técnica surge siempre que el análisis se centre en el "contenido" *manifiesto* y *latente* de la información "verbal" (extraída de distintas fuentes: documentos, prensa escrita, transcripciones de entrevistas, grabaciones radiofónicas, de programas de televisión) o "visual" (de la observación de imágenes estáticas –fotografías, cuadros– y/o en movimiento –una grabación en vídeo, por ejemplo).

Pero, ¿en qué consiste el *análisis de contenido?,* ¿qué peculiaridades distinguen a esta variedad de análisis?, ¿cuáles son sus límites y posibilidades? A estas y otras preguntas tratará de darse respuesta en las páginas siguientes.

10.1. El análisis de contenido como técnica de análisis

El *análisis de contenido* se perfiló, en sus orígenes, como una técnica analítica eminentemente *cuantitativa.* Berelson (1952) la definió como la descripción *objetiva, sistemática* y *cuantitativa* del *contenido manifiesto* de la comunicación. No obstante, en la actualidad se reconoce la pluralidad analítica existente, de hecho, bajo la rúbrica

común del "análisis de contenido". Éste no se limita a la "cuantificación" del *contenido manifiesto* de la comunicación; sino que también aborda la interpretación del *contenido latente*.

Tesch (1992) diferencia dos tipos genéricos de *análisis de contenido:* el *clásico* y el *etnográfico*. El primero se ajusta a la definición de Berelson (1952), mientras que el segundo (el *etnográfico),* responde a una redefinición desde posturas *cualitativas* (Altheide, 1987), que pone el acento en el *análisis reflexivo de documentos*. Este segundo tipo se dirige, no tanto a la *descripción,* como a la *comprensión* de los *significados latentes,* y a la *verificación* de relaciones teóricas.

Dada la especificidad del presente manual (sobre la *metodología cuantitativa),* la exposición que aquí se hace del *análisis de contenido* se ciñe a la modalidad *clásica* o *cuantitativa*. Para su complementación con el *análisis de contenido etnográfico,* remito al manual de metodología cualitativa de esta misma colección, y a cualquier texto especializado en el *análisis cualitativo*.

10.1.1. Características esenciales del análisis de contenido cuantitativo

Los tres rasgos definitorios del *análisis de contenido cuantitativo* están ya presentes en la definición clásica de Berelson (1952): *objetividad, sistematicidad* y *cuantificación* de los *contenidos manifiestos* de la comunicación. A éstos se suma otro fundamental: la *inferencia* de los "datos" al "contexto" de referencia. De hecho, éste es –siguiendo a Krippendorff (1980, 1990)– el propósito primordial del análisis: realizar, a partir de ciertos datos, *inferencias* válidas y replicables, que puedan aplicarse a su *contexto*. Ello exige que el análisis reúna condiciones mínimas de *validez* y de *fiabilidad,* en el proceso de medición.

En el *análisis de contenido cuantitativo* el énfasis no recae en los aspectos semánticos o sintácticos de los textos, sino en la "cuantificación" de sus integrantes (palabras, expresiones, frases, temas); es decir, en la medición de su "frecuencia" de aparición en el texto.

Esta "cuantificación" permite, precisamente, la comparación de los "contenidos" de diferentes documentos y, sobre todo, la *inferencia*. Ésta sería –de acuerdo con Weber (1994)– difícilmente alcanzable por otros medios. Pero, para que pueda, a partir de los *contenidos "manifiestos",* inferirse los *contenidos "latentes"* (no explícitos en el documento, y relativos tanto al mensaje, como a sus agentes emisores y receptores), la "cuantificación" ha de hacerse de forma *sistemática* y *objetiva*.

La *objetividad* se halla relacionada con el uso de procedimientos "rigurosos y replicables" (Scott, 1990: 130). No sólo se han de elegir *indicadores* que constituyan una representación "válida" de los *conceptos*. También ha de lograrse la coincidencia en la *codificación* de un mismo documento, por personas diferentes, si se pretende que la clasificación efectuada se considere "fiable".

En suma, el *análisis de contenido cuantitativo* puede caracterizarse como un análisis no limitado a la *descripción,* sino orientado a la *inferencia*. La *descripción* (o enu-

meración de las características del texto) constituye –en conformidad con Bardin (1986)– la primera etapa del análisis; la *interpretación* (la significación acordada de estas características), la última. Entre ambas se encuentra la *inferencia,* como un procedimiento intermedio, que permite el paso de la *descripción* a la *interpretación.*

10.1.2. Fases de un análisis de contenido

Bardin (1986) diferencia tres momentos cronológicos fundamentales en la organización del *análisis de contenido:*

a) El *preanálisis* o fase de organización, de *operacionalización* de las ideas originarias de la investigación. Incluye:

1) La especificación de los *objetivos* de la investigación y la formulación de las *hipótesis.*
2) Elaboración de los *indicadores* en que se apoyará la interpretación de los datos. Éstos se obtendrán a partir de las *definiciones operativas* de los *conceptos teóricos* básicos en la investigación.
3) Elección de los *documentos* a analizar: la *muestra* del estudio.

b) Explotación del material. Incluye:

1) Elección de las *unidades de codificación* (palabras, expresiones, frases, temas, imágenes...; depende de qué se analice). De ellos, se cuantificará la *frecuencia,* e *intensidad,* de aparición en el documento. También pueden efectuarse enumeraciones de carácter temporal (cuándo) y espacial (dónde aparecen).

   Además de las *unidades de codificación,* se identifican las *unidades de contexto,* siguiendo la recomendación de Holsti (1969). Una misma palabra, por ejemplo, puede adoptar significados diferentes, dependiendo del *contexto* donde se ubique. Razón por la que se aconseja su contextualización. Ésta ayudará en la *interpretación* del significado de las *unidades de codificación.*
2) Determinación de las *categorías* a emplear en la *codificación* del "contenido" de los documentos, a partir de las *hipótesis* de la investigación.

   Esta fase es exigida, cuando se realiza un *análisis de contenido categorial* o temático. Si bien, ésta no es la única variedad de *análisis de contenido,* aunque sea la más habitual.
3) El *registro* y *cuantificación* de los *datos,* de acuerdo con los *indicadores* elegidos.

c) Tratamiento e interpretación de los resultados. A los datos reunidos se les da tratamiento estadístico, a semejanza de lo descrito en el Capítulo 9. Se comienza con el *análisis univariable.* Si las características de los datos lo permiten, puede concluirse con la aplicación de alguna técnica de análisis multivariable.

Los resultados se someten a *pruebas estadísticas* y a *tests de validez.* En caso afirmativo, el investigador podrá hacer *inferencias* e *interpretaciones* de los

resultados de la investigación. Estas *interpretaciones* pueden actuar como punto de partida en una nueva investigación.

En la Figura 10.1 se resumen las fases principales en un *análisis de contenido cuantitativo*.

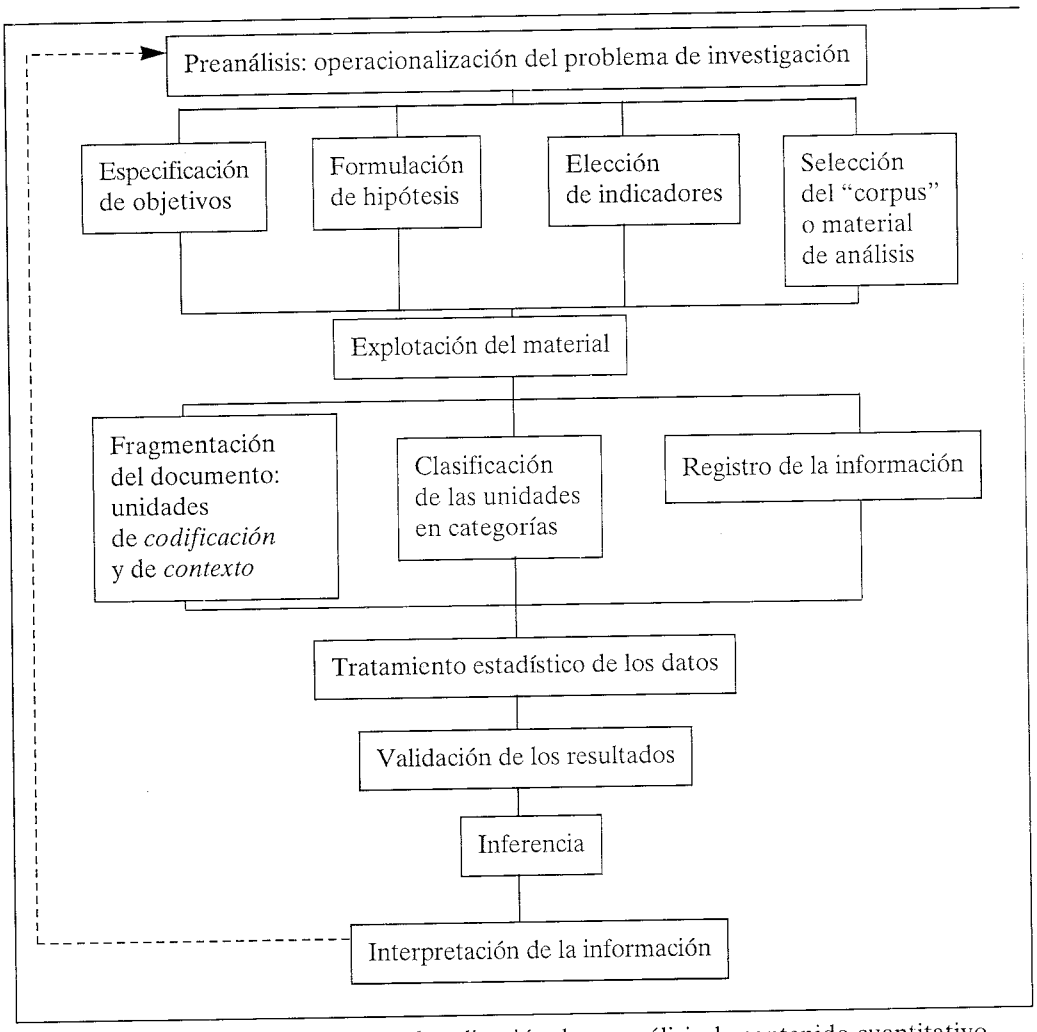


Figura 10.1. Fases principales en la relización de un análisis de contenido cuantitativo.

De las fases representadas en el gráfico, únicamente se comentarán, a continuación, los aspectos más específicos del *análisis de contenido*. Sobre los demás, remito a los capítulos iniciales del libro.

10.2. La selección del "corpus" de documentos o material de análisis

En función de los objetivos de la investigación, se eligen las *unidades de análisis* y la *estrategia* a seguir en la recogida de información.

Para efectuar un *análisis de contenido,* previamente ha de delimitarse el *"corpus"* o material a analizar. Esta delimitación incluye tanto a la *población* de documentos, como la concreción del procedimiento de *muestreo* a seguir en su selección.

Salvo que el *universo* sea de escasa amplitud, lo habitual es proceder a la extracción de una *muestra* representativa de dicho *universo*. En esta selección se siguen los mismos procedimientos muestrales descritos en el Capítulo 5. Concretamente, se consideran las características de la población de documentos (escritos, visuales, etc.), y el tipo de *inferencia* que desea realizarse. En general, se recomienda seguir procedimientos de selección *aleatorios*. Con estos se logra una mayor representatividad de la *muestra*.

La *selección muestral* comienza con la elección de la *fuente de comunicación* (periódicos, revistas, libros, canciones, imágenes). Prosigue con la extracción de *documentos* y, en caso de que éstos sean muy extensos, de *partes* de estos documentos.

EJEMPLO DE UNA SELECCIÓN MUESTRAL PARA UN ANÁLISIS DE CONTENIDO

Imaginemos que se desea analizar el tratamiento dado a los temas de corrupción en la prensa española. Uno de los posibles *diseños muestrales* incluiría las *fases* siguientes:

a) *Elección de fuentes.* Habría que decidir qué periódicos representarán el "*corpus*" de la investigación. El investigador puede optar entre elegirlos "intencionalmente", o seleccionarlos de forma *aleatoria,* siguiendo cualquiera de los procedimientos expuestos en el Capítulo 5 (desde el *muestreo aleatorio simple,* hasta el de *conglomerados).*

   Si se busca una *muestra* representativa de la pluralidad de periódicos nacionales (de ámbito estatal, autonómico, o provincial), puede hacerse un *muestreo aleatorio estratificado* por Comunidades autónomas e ideología de las editoriales de los periódicos.

b) Seleccionados los periódicos, se procede a la *elección,* en cada uno de ellos, de unos *números* concretos. En esta segunda fase del *diseño muestral,* la selección atiende, habitualmente, a "fechas" de publicación. Como indican Sánchez Carrión (1985) y Weber (1985, 1994), los "contenidos" de los periódicos difieren por día de la semana (laboral o festivo) y período del año (diferenciando, sobre todo, el período vacacional del resto del año). Estos y otros aspectos deberán tenerse presentes en la selección (intencional o aleatoria) de las *fechas* de edición de los periódicos que forman la *muestra.*

c) *Extracción de textos en los documentos de la muestra,* si las dimensiones de éstos lo exigen. En este último estadio del *diseño muestral,* la selección suele ser, a diferencia de las anteriores, *intencional.* Se buscan, de forma intencional, "aquellas unidades que tratan sobre el tema objeto de estudio" (Sánchez Carrión, 1985: 96).

Weber (1994) recomienda el análisis del texto completo, siempre que sea posible. Ello ayudará a preservar la coherencia semántica del texto como unidad. Cuando no sea factible el análisis del documento completo, el investigador deberá considerar la *estructura del texto* en la *selección muestral.* En concreto:

1) Podrían excluirse las secciones de *introducción* y de *cierre.*
2) Cada *muestra* debería constar de uno o más *párrafos* enteros para no alterar la coherencia semántica.
3) Las *frases* no deberían muestrearse. Analizar frases separadas (incluso aquellas que se han extraído de un mismo tema), destruye la coherencia semántica. Esto dificulta la *validación* posterior e *interpretación* de los resultados del *análisis de contenido.*

Por último, respecto al *tamaño de la muestra,* éste dependerá, principalmente, de la variabilidad de los textos dentro de un mismo periódico (Sánchez Carrión, 1985). Cuánto más homogéneo sea su *contenido,* menor *tamaño muestral* se precisará para representar a la *población.*

10.3. La explotación del material

A la selección de documentos *(población, muestra)* sigue la determinación de criterios para la extracción de la información en ellos "contenida". Los *criterios* varían dependiendo, fundamentalmente, de los objetivos de la investigación y de las características de los documentos. No obstante, pueden señalarse unas *pautas comunes* en los *análisis de contenido cuantitativo.* Éstas conciernen a la creación del *esquema de codificación* y la preparación del documento para su posterior tratamiento estadístico. En concreto, se diferencian tres estadios básicos en esta fase esencial al *análisis de contenido:*

a) La fragmentación del documento en *unidades de codificación* y de *contexto.*
b) La clasificación de las *unidades* en *categorías.*
c) El *registro de la información.*

10.3.1. La fragmentación del documento en unidades de codificación y de contexto

Indistintamente de cómo se vaya a analizar la información, una vez elegidos los documentos, éstos han de fragmentarse en *unidades* que ayuden a su organización y análisis.

Las *unidades* se dividen en *unidades de codificación* (o *registro)* y las *unidades de contexto.* Las *unidades de codificación* (o *registro)* se definen como "la mínima porción de contenido que el investigador aísla y separa por aparecer allí uno de los símbolos,

palabras, *slogans* o temas que el investigador considera significativos" (López Aranguren, 1989: 394). Corresponden, por tanto, "al segmento de contenido que será necesario considerar como unidad de base con miras a la categorización y al recuento frecuencial" (Bardin, 1986: 79).

Las *unidades de codificación* que se emplean con mayor frecuencia son las siguientes:

a) La *palabra*. Ésta constituye la *unidad de codificación* más básica. A ello contribuye su fácil identificación. Si bien, como indica Weber (1994), ha de tenerse presente el problema de los distintos sentidos que puede tener una misma palabra, dependiendo del *contexto* en el que se emplea. No todos los paquetes informáticos disponen de rutinas de desambiguación o de contextualización de palabras específicas.

La *codificación* de *palabras* puede incluir a todas las comprendidas en el texto, o restringirse, por el contrario, a aquéllas consideradas "claves", en conformidad con los objetivos de la investigación.

b) La *frase,* cuando se está interesado en la cuantificación de frases que contienen un mismo significado (por ejemplo, "a favor" o "en contra" de un determinado tema).

Si la *frase* es extensa o compleja, puede decidirse su división en unidades más pequeñas.

c) El *tema* en torno al cual gira el discurso del documento. Holsti (1969) lo define como la unidad textual que hace referencia a uno de los siguientes aspectos: quién percibe; el agente de la acción; la misma acción; y el objeto de la acción.

Por ejemplo, el *tema* representa una *unidad de codificación* habitual en estudios de motivaciones, opiniones, actitudes, valores, o creencias (Bardin, 1986).

d) El *párrafo*. La *codificación* de *párrafos* enteros es practicada, sobre todo, en el *análisis de contenido manual,* cuando se carece de la ayuda de programas informáticos al efecto.

Weber (1994: 265) descarta esta cuarta *unidad de codificación* por su baja fiabilidad. La experiencia de este autor le lleva a afirmar que resulta "más difícil alcanzar una fiabilidad elevada cuando se codifican unidades grandes (como párrafos), que cuando se codifican unidades más pequeñas como palabras".

e) El *texto entero,* si éste puede codificarse en su globalidad. Ello exige que el *texto* sea *breve* (titulares de periódicos, editoriales, o las respuestas a una *pregunta abierta,* por ejemplo). En caso contrario, será "difícil alcanzar una elevada fiabilidad cuando se codifican textos completos" (Weber, 1994: 265). A ello se suma otra condición importante: la idea dominante o principal ha de ser suficiente para el objetivo buscado (Bardin, 1986: 81).

f) El *personaje* o actor al que se le atribuyen las acciones que figuran en los documentos analizados.

Ésta constituye una *unidad de codificación* adecuada "siempre que el texto narre una acción" (Sánchez Carrión, 1985: 97). Bardin (1986) cita como ejemplos, las obras de ficción (films, emisiones, novelas...), que pueden analizarse siguiendo a sus personajes.

De estas *unidades de codificación,* el investigador escogerá una o varias. La elección dependerá de la conjunción de tres *aspectos* básicos:

a) Los objetivos de la investigación.
b) Las características del documento.
c) Los medios disponibles para la realización del *análisis de contenido* (manual o mediante ordenador).

Además de las *unidades de codificación,* en los documentos también se diferencian las *unidades de contexto.* Éstas son de gran utilidad para la comprensión de las *unidades de codificación.*

Bardin (1986: 82) define las *unidades de contexto* como "el segmento del mensaje cuyo tamaño (superior a la unidad de registro) es óptimo para captar la significación exacta de la unidad de registro". Si la *unidad de registro* fuera, por ejemplo, la *palabra,* la *unidad de contexto* sería la *frase;* del *tema,* el *párrafo;* del *párrafo,* el *documento entero.*

La *unidad de contexto* hace, por tanto, referencia al "contexto" donde se ubica la *unidad de codificación.* Su identificación resulta, en consecuencia, básica para la comprensión del *significado* de la *palabra,* la *frase,* o cualquiera que sea la *unidad de codificación* elegida.

10.3.2. La clasificación de las unidades en categorías

En el *análisis de contenido "categorial",* las *unidades de registro* se codifican en *categorías,* que singularicen su "contenido". Estas *categorías* proporcionan una denominación genérica a unidades de contenido similar, contribuyendo, de esta forma, a su clasificación. Para ello, las *categorías* han de cumplir los tres *requisitos* –referidos en el Capítulo 4, apartado 4.2– comunes en la medición de variables: *exhaustividad, exclusividad* y *precisión.*

No sólo han de cubrir toda la variedad del objeto de estudio. También han de evitar la "ambigüedad" en el proceso de medición, si se quiere que el *análisis de contenido* reúna condiciones de *validez* y de *fiabilidad.* Quiere esto decir que cada *unidad de registro* sólo puede clasificarse en una única *categoría.*

A esos tres *requisitos* básicos se añaden otros dos fundamentales:

a) La *pertinencia* o adecuación de las *categorías* a los *objetivos* de la investigación.
b) La *homogeneidad:* la creación de *categorías* ha de responder a un único *principio de clasificación.*

López Aranguren (1989: 398-401) diferencia los siguientes *principios de clasificación*. De ellos derivan la mayoría de las *categorías* empleadas en el análisis de material escrito:

a) *Materia, tema* o *asunto:* ¿de qué trata la comunicación?
b) *Dirección* o *punto de vista:* el tratamiento o presentación (favorable-desfavorable-neutral; positiva-negativa) que la comunicación hace de la materia o asunto del que trata.
c) *Criterio* o *valor:* sobre qué se basa la clasificación en términos de "dirección" o punto de vista. Por ejemplo, "moralidad-inmoralidad", "fuerza-debilidad".
d) *Autoridad, fuente* o *prueba:* personas, grupos, instituciones u objetos que en la comunicación se citan en apoyo de una afirmación, o en cuyo nombre se hacen declaraciones o afirmaciones.
e) *Objetivo* o *meta* que se persigue.
f) *Medio* o *método* para alcanzar los objetivos.
g) *Actores:* personas, grupos o instituciones que en el texto se representan como realizadores de determinados actos. Por ejemplo, países, grupos religiosos, económicos, etc.
h) *Características personales* y *rasgos de carácter:* edad, sexo, optimismo, depresión, entre otras.
i) *Origen:* lugares de origen de las comunicaciones.
j) *Destino,* cuando la comunicación no se dirige al público en general, sino a grupos específicos (como estudiantes, jubilados, mujeres, por ejemplo).
k) *Forma de la afirmación:* hechos, expectativas, preferencias, demandas, ...
l) *Intensidad,* cuando se clasifica el componente emocional o sentimental de las clasificaciones, el entusiasmo, la agitación o la excitación.
m) *Mecanismo* o *recurso estilístico* o *retórico,* que se emplea con fines propagandísticos, persuasivos o de tendencias.

EJEMPLO DE CLASIFICACIÓN DE UNIDADES EN CATEGORÍAS

Como ilustración de lo expuesto, a continuación se va a exponer el *sistema de categorías* desarrollado en una investigación real, en la que se aplica el *análisis de contenido*. Se trata del estudio llevado a cabo por M. Clemente y M. A. Vidal, publicado en 1994 con el título "La violencia simbólica: la televisión como medio generador de la delincuencia" (en la revista *Apuntes de Psicología,* n.º 41-42, pp. 47-60).

En esta investigación, los autores organizaron la información extraída de series de dibujos animados –emitidos en cadenas públicas (TVE1, TVE2, Telemadrid) y privadas (Tele 5, Antena 3 y Canal Plus)– en un *sistema de categorías* diseñado al efecto. Este *sistema* se confecciona conforme a tres *criterios de clasificación*:

a) Quién realiza el acto agresivo.
b) Sobre quién se comete el acto agresivo.
c) De qué acto agresivo se trata.

En los dos criterios de clasificación primeros, los autores diferencian las mismas siete *agrupaciones de categorías:*

a. Sexo
1. Varón
2. Mujer
3. Sin determinar

b. Edad
1. Niños
2. Adolescentes
3. Adultos
4. Sin especificar

c. Apariencia
1. Humano
2. Animal
3. Planta
4. Objeto

d. Capacidad de decisión individual
1. La tiene
2. No la tiene

e. Individual o grupal
1. Uno sólo
2. Grupo
3. No está

f. Rol
1. Protagonista
2. Habitual
3. Otros

g. Calificación del rol
1. Bueno
2. Malo
3. Sin especificar

En cambio, en la especificación de "qué acto agresivo se realiza", los autores distinguen tres *agrupaciones de categorías:*

a. Tipos de agresiones
1. No verbales
2. Verbales leves o actos irónicos
3. Verbales graves o fuertes
4. Amenazas
5. Agresiones físicas leves (una sola)
6. Agresiones físicas fuertes (más de una)
7. Agresiones con objetos no destinados a agredir
8. Armas blancas
9. Armas

b. Grado de proactividad del acto
1. Proactivo
2. Reactivo

c. Grado de daño manifiesto
1. No aparece
2. Leve
3. Fuerte
4. Muerte

La información analizada corresponde a series de dibujos animados emitidos desde el 29 de noviembre al 5 de diciembre de 1993, ambos días inclusive (tanto en días de diario, como en fines de semana). En conformidad con el *sistema de categorías* señalado, la información se cuantificó en forma de *frecuencia (absoluta* y *porcentual)* de aparición. A modo de ejemplo, en el Cuadro 10.A se recoge la tabulación efectuada para los primeros cinco *sistemas de categorías.*

CUADRO 10.A. Distribuciones de frecuencias para cinco sistemas de categorías.

		Características del agresor		*Características del agredido*	
Categoría		*Frecuencia*	*Porcentaje*	*Frecuencia*	*Porcentaje*
Sexo	1	382	64	257	43
	2	86	14	168	28
	3	131	22	174	29
Edad	1	30	4	83	14
	2	76	11	82	14
	3	271	51	230	38
	4	291	44	206	34
Apariencia	1	287	49	226	38
	2	192	33	235	39
	3	28	5	23	4
	4	79	13	111	19
Capacidad decisión	1	461	78	387	79
	2	132	22	106	21
Grupo	1	494	83	410	68
	2	93	16	190	31
	3	4	1	8	1

Una vez diseñado el *esquema de codificación,* se recomienda (véase Bardin, 1986; o Weber, 1994) su comprobación en una pequeña *muestra* (de la misma población de documentos), previa a su utilización definitiva para el registro de la información. Esta precomprobación ayudará a precisar la adecuación del *esquema de codificación* elaborado para la consecución de los objetivos de la investigación (la clasificación de las *unidades de registro,* de *contexto,* y el *sistema de categorías* creado).

10.3.3. El uso de programas informáticos para el análisis textual

Lo hasta ahora expuesto constituye el procedimiento usual, cuando el *análisis de contenido* se realiza manualmente. Sin embargo, como ya señalaba Mochmann en 1985, cada vez van adquiriendo mayor protagonismo los "procedimientos de *clasificación empírica*". En ellos, el investigador no define, *a priori,* los *conceptos teóricos.* Por el contrario, éstos resultan de la realización de procedimientos estadísticos (como el *análisis factorial* o el *análisis de conglomerados),* a partir de las *matrices de correlaciones* de palabras incluidas en los textos.

Como ilustración de este proceder, se hace mención del programa WORDS, diseñado con la finalidad de excluir cualquier influencia del investigador en la *categorización* de los textos. A tal fin, se siguen las *fases* siguientes:

a) El documento se divide en *segmentos* (páginas, párrafos, entrevistas, por ejemplo).
b) Se suprimen las *palabras funcionales* (artículos, conjunciones).
c) Se determina la *frecuencia* de aparición de cada *palabra* en cada *segmento*.
d) Se selecciona un *subconjunto* de esas *palabras* para su análisis.
e) Se calcula una *matriz de correlaciones* de las palabras incluidas en ese *subconjunto*.
f) A partir de la *matriz* se realiza un *análisis factorial*. De él se obtendrá un número de *factores,* que corresponderán a los principales *temas* de *contenido* del documento analizado (Mochmann, 1985: 37).

Asimismo, la mayoría de los programas informáticos para *análisis de contenido* no sólo proporcionan *listados de palabras,* acompañados con su *frecuencia* de aparición en el texto. También pueden proporcionar el *contexto* que arropa a la *unidad de codificación*. Esto adquiere especial relevancia en casos de ambigüedad, cuando la *palabra* presenta distintas connotaciones en *contextos* diferentes. Para su correcta clasificación en *categorías,* se crean, entonces, *índices* o *diccionarios generales*. Éstos actúan a modo de *thesaurus,* de diccionarios, que recogen los posibles sentidos de las *palabras,* en función del *contexto* en que se hallen.

El *General Inquirer* (el primero de los programas para *análisis de contenido,* desarrollado inicialmente por Stone y colaboradores, en 1966), por ejemplo, incluye varios *diccionarios generales,* especializados en áreas de conocimiento. Entre ellos se encuentran los siguientes: "Harvard Psychological Dictionary", "Stanford Political Dictionary", "Need Achievement Dictionary", "Santa Fe Anthropological Dictionary", o "Laswell Value Dictionary" (LVD).

Estos *diccionarios* proporcionan al investigador numerosas *categorías* (generalmente de 60 a 150, e incluso más), así como definiciones o reglas para la correcta clasificación de *palabras* en *categorías*. A tal fin, se considera el sentido que adopte en el *contexto* en que aparece.

Las *categorías* se representan mediante abreviaciones o *etiquetas,* que figuran en tablas. El número al final de la *palabra* indica el número de sentidos particulares de las palabras. De esta forma, la creación de *diccionarios generales* ayuda a la sistematización y estandarización del procedimiento de clasificación de las *palabras* en *categorías*.

EJEMPLO DE DICCIONARIO GENERAL

La presente ilustración se ha tomado de la obra de Bardin (1986: 100-101). Corresponde al *Stanford Political Dictionary* (de Holsti), para el análisis de documentos políticos. En él pue-

den registrarse hasta un total de 4.000 palabras, aproximadamente, según seis o cuatro conceptos clave. Por ejemplo:

Conceptos clave	*Número de palabras*	*Ejemplos de palabras*
Afecto positivo	977	Mutuo, natural, normal, puro.
Afecto negativo	1.513	Escándalo, profano, rehusar, repugnante.
Fuerza	1.391	Acero, piedra, espada, duro.
Debilidad	579	Delgado, ignorante, debajo.
Actividad	1.218	Reacción, reino, contacto, viaje.
Pasividad	722	Inmutable, espera.
Sobreestimación	128	Absolutamente, exactamente, siempre.
Subestimación	50	Aunque, aparentemente, aproximadamente
Negación	6	Diferente, ni uno ni otro.

Además de los *diccionarios,* la mayoría de los programas incluyen el procedimiento KWIC *(Key Words In Context),* que relaciona las *palabras* con el *contexto.* Por ejemplo, el programa TEXTPACK (de Mohler y Zuell, 1990, ZUMA, reléase 4.0) incluye seis operaciones básicas, entre las cuales se incluye el procedimiento KWIC y la elaboración de *índices* (o *diccionarios).* Las operaciones que incluye este paquete son:

a) *Frecuencia de palabras* diferentes para uno o más textos (procedimiento FREQ).
b) *Índice:* un listado de palabras y sus referencias en el texto (procedimiento XREFF).
c) *Palabras claves en su contexto (concordancia),* mediante el procedimiento KWIC.
d) *Comparación de vocabulario,* mediante el procedimiento WORDCOMP *(Word Comparison).* Este procedimiento está diseñado para la comparación de palabras con una denominación común o diferente.
e) *Selección de textos utilizando información externa* al texto, en forma de fichero numérico (un fichero SPSS, por ejemplo).

En estos casos puede aplicarse el programa SUBSEL. Éste permite la selección de un texto, que corresponde a una combinación específica o valores en un fichero numérico.

f) *Codificar/categorizar palabras* que figuran en secciones concretas del texto, y su posterior inclusión en un fichero numérico (TAGCODER).

A partir de esta información podrán después efectuarse análisis estadísticos (mediante paquetes estadísticos como el SPSS, BMDP o SAS, por ejem-

plo). Para ello se utilizarán no *palabras*, sino *grupos de palabras* clasificadas en *categorías,* en función de su significado común.

Otros programas que tan sólo se mencionan aquí son: EVA (creado para el análisis de los titulares de los diarios), ANACONDA (para la codificación de las *preguntas abiertas* de los *cuestionarios),* SPENCE (para el análisis de los *protocolos* de las *entrevistas* psiquiátricas), y el QUESTER (dirigido al análisis del *contexto)* (Mochmann, 1985).

A estos programas hay que sumar los desarrollados en fechas más recientes. Como son el LEXICLOUD (por el Laboratorio de Lexicometría y Textos Políticos de la Escuela Normal Superior de Fontenay-St.Cloud, para el establecimiento de relaciones estadísticas entre las unidades léxicas), y el sistema ALCESTE (creado por Reinert en 1990, para el análisis léxico por contexto de un conjunto de segmentos de texto). El detalle de estos programas puede encontrarse en el artículo de Julia Behar (1993).

La ayuda de estos u otros programas facilita, considerablemente, la práctica del *análisis de contenido*. Entre sus *ventajas* destacan las siguientes:

a) Pueden analizarse, simultáneamente, textos complejos, y de considerable magnitud, en un breve período de tiempo.
b) Se introduce una mayor estandarización y sistematización en el análisis, que mediante procedimientos manuales. Ello facilita la comparabilidad de los resultados analíticos.
c) Los errores de *codificación* se reducen considerablemente, al no depender ésta de la actuación del equipo investigador.

A estas *ventajas* hay que añadir la reducción en costes (económicos y temporales) que supone la aplicación de *scanners* (con reconocimiento de caracteres) para la grabación óptica de textos en el ordenador. Atrás ha quedado el tedioso procedimiento primigenio de transformación del texto en *tarjetas perforadas,* para su posterior lectura por el programa informático. Además, el uso de *scanners* presenta otra ventaja importante: la reducción de los *errores de grabación*. Los textos se introducen en el ordenador tal como aparecen en el original.

10.4. El tratamiento estadístico y la interpretación de los resultados

En el *análisis de contenido cuantitativo* se aplican las mismas *técnicas estadísticas* descritas en el Capítulo 9. A la *segmentación* de los textos sigue el *listado de frecuencias* de aparición de *palabras, frases,* o cualquiera que sea la *unidad de codificación* escogida en la investigación.

En este estadio inicial (a veces único), las *listas KWIC de palabras claves* son de gran utilidad. Estas listas muestran el *contexto* más inmediato de la *palabra,* lo que ayudará en la interpretación de su significado.

Las palabras que únicamente sirven para articular el texto (artículos, pronombres, conjunciones, preposiciones) se excluyen de los análisis. El interés, en cambio, se dirige a la cuantificación de palabras "claves"; aquellas que conforman el léxico del discurso (sustantivos, adjetivos, verbos, adverbios).

Pero, como reconoce Weber (1994: 294):

> "Aunque las listas de frecuencias de palabras revelan cambios o diferencias en el énfasis entre los documentos, deben utilizarse con precaución. Las frecuencias de palabras no revelan mucho sobre la asociación entre las palabras."

Cuando el interés del investigador va más allá de la mera descripción frecuencial de las unidades de codificación por separado, puede aplicarse cualquiera de las *técnicas multivariables* descritas en el Capítulo 9. Lo que normalmente exige la previa codificación numérica del *contenido manifiesto* del material verbal.

Una de las *técnicas multivariables* más aplicadas en el *análisis de contenido* es el *análisis factorial.* En éste, las *variables observadas* suelen ser ahora las *categorías* diferenciadas. Las *variables latentes,* por el contrario, hacen referencia a *temas* genéricos, "latentes" en el texto e, igualmente, deducidos de las *variables observadas* que conforman cada *factor* obtenido. De este modo, como en cualquier *análisis factorial,* las *categorías* que correlacionan con un mismo *factor* (es decir, aquellas que presenten un *"factor loading"* superior a .30) se interpretarán como representando un mismo *tema* en el texto analizado.

Los resultados numéricos se *validan* e *interpretan* en relación con el *tema* al que hacen referencia. Los resultados pueden corroborar las *hipótesis* de la investigación o, por el contrario, contradecirlas. Lo que puede provocar la realización de una nueva investigación.

Como ilustración de lo aquí expuesto, se han seleccionado dos investigaciones reales, en las cuales el *análisis de contenido cuantitativo* se ha materializado de forma distinta. La primera investigación ejemplifica la aplicación del *análisis factorial;* mientras que en la segunda, el análisis se restringe al recuento de *frecuencias.*

10.4.1. Ejemplos desarrollados de análisis de contenido en dos investigaciones reales

- Cabrera, J. (1992): *La nación como discurso. El caso gallego,* Madrid. CIS. *Monografía,* n.º 126.

En esta investigación se analizan distintas manifestaciones del sistema ideológico nacionalista:

a) El nivel *léxico* y *semántico,* que constituyen las estructuras superficiales del campo de la manifestación.
b) El nivel *semiológico,* que trata de desentrañar la significación profunda de dicho sistema.

Para cada nivel se ha optado por una *técnica de análisis* diferente, combinándose técnicas *cuantitativas (lexicometría)* y *cualitativas* (análisis *semántico-categorial* y análisis *semiológico).*

El *corpus* que se somete al *análisis lexicométrico* (del vocabulario que conforma el discurso) está integrado por las respuestas obtenidas mediante *entrevista abierta.* El *guión* de la *entrevista* comprende tres aspectos concretos:

a) La biografía de cada entrevistado.
b) El proceso de socialización nacionalista.
c) La posición teórico-nacionalista de los entrevistados.

La *palabra* se toma como *unidad de medida,* procediéndose a su "lematización", que comprende:

a) La *reagrupación de las formas heterogéneas de un mismo vocablo.* Consistió en reducir las formas plurales a singulares, los femeninos a masculinos, y las diversas formas de un verbo a su expresión en "infinitivo" (a excepción de los "participios". Éstos se analizaron contextualizadamente, con el propósito de determinar si constituían un adjetivo).
b) La *separación de las formas homógrafas que muestran vocablos diferentes.* Se analizaron formas con igual grafía que hacían referencia a contenidos diferentes. Estos *homógrafos* fueron identificados con un dígito, o un *dígito* y una letra, a modo de *sufijos.*

Por *ejemplo,* la forma adjetiva de "gallego" se distinguió de sus formas sustantivas, quedándose como sigue:

- "Gallego" = adjetivo
- "Gallego2" = sustantivo
- "Gallego2a" = sustantivo referido al idioma

En general, se adoptaron las claves siguientes: "2" para sustantivos; "2a" para sustantivos referidos a idiomas; "3" para adjetivos; "4" para adverbios; y, "5" para aquellos nombres propios de personas o ciudades que podían inducir a confusión.

Finalizada la *lematización,* se efectúa el recuento mecánico de los textos. Para ello se diferencia, a nivel *léxico* (de vocabulario), dos tipos diferentes de *palabras:*

a) *Palabras de relación o gramaticales,* aquellas que por sí mismas carecen de significado: artículos, pronombres, preposiciones y conjunciones.
b) *Palabras lexicales* o formas plenas de significado. Éstas conforman el *léxico* del discurso: sustantivos, adjetivos, adverbios y verbos.

La investigación se centra en el "discurso lexical", en el análisis estadístico de los rasgos de contenido. En el Cuadro 10.2 figura la distribución general de frecuencias para cada uno de los 12 sujetos cuyos discursos forman la *muestra* del estudio. Los diversos "subcorpus" se distribuyen en función de la extensión del:

a) *Número de ocurrencias* (N): el número de ítems. Refleja, por tanto, su extensión.
b) *Vocabulario* (V): número de formas diferentes empleadas.
c) *Hapax:* aquellas formas cuya frecuencia de aparición es igual a 1.
d) *Número de ocurrencias funcionales* ($FF_{(n)}$): items que se corresponden con el vocabulario funcional ($FF_{(v)}$).
e) *Número de ocurrencias lexicales* ($FL_{(n)}$), con la extensión del vocabulario correspondiente.

Cada uno de estos valores figura acompañado de su correspondiente *valor de rango* (r) para cada columna. El *corpus* total (suma de los diversos *subcorpus* correspondientes a cada emisor) tiene una extensión de 133.249 palabras. Éstas se distribuyeron porcentualmente en los diversos *subcorpus* como se refleja en el Cuadro 10.1.

CUADRO 10.1. Distribución general de frecuencias.

	N	r	V	r	Hapax	r	$FF_{(n)}$	r	$FF_{(v)}$	r	$FL_{(n)}$	r	$FL_{(v)}$	r
GM	5524	1	1113	1	613	2	3066	1	162	2	2458	1	951	1
VA	6521	2	1151	2	598	1	3766	2	160	1	2755	2	991	2
FE	8378	3	1601	7	875	8	4624	3	175	5	3754	3	1426	7
GS	9593	4	1479	3	762	4	5543	5	183	8	4050	5	1296	3
JB	9605	5	1580	5	821	7	5314	4	177	6	4291	6	1403	5
CH	9763	6	1503	4	771	5	5722	6	171	3	4041	4	1332	4
PA	11802	7	1581	6	698	3	6618	8	174	4	5184	7	1407	6
CC	12098	8	1869	10	959	10	6556	7	189	9	5542	10	1680	10
LS	12655	9	1751	9	924	9	7218	9	190	10	5437	8	1561	9
RA	13359	10	1695	8	792	6	7845	10	178	7	5514	9	1481	8
LF	16794	11	2206	11	1136	11	9166	11	198	12	7628	11	2008	11
BA	17157	12	2284	12	1158	12	9492	12	192	11	7665	12	2092	12
Total	133249				1010		74930				58319			
							%N: 56,23				%N 43,77			

En el Cuadro 10.2 destacan los discursos de los sujetos BA y LF, cuya extensión supera a la del resto de los entrevistados. Ambos acumulan algo más de la cuarta parte de todo el *corpus*. Los discursos de RA, LS, CC y PA forman un segundo bloque. Sus discursos superan el millar de *ítems,* pero no alcanzan a superar el millar quinientos mil. Por lo que, acumulan el 37,5% del total.

Respecto a la *lexicalidad,* en esta investigación se confirma una norma observada en la generalidad de los *análisis de contenido cuantitativo:* las *formas funcionales* (FF) superan el 50% de la extensión general. En este caso, el 56%. En la investigación se analizará, en consecuencia, el 44% restante que conforma el "discurso lexical": las formas que caracterizan el discurso sobre la nación.

A *nivel léxico* se diferencia entre:

a) *Vocabulario original:* toda forma que sólo se encuentra en un emisor de la *muestra.*
b) *Vocabulario común:* formas que tienen al menos una ocurrencia en todos los emisores.
c) *Vocabulario de base:* formas que tienen una distribución regular en todos los emisores. Por lo que no introducen diferencias.
d) *Vocabulario específico:* el que caracteriza a cada uno de los emisores.

El subconjunto de formas así obtenido constituye el *corpus* sobre el que se realizan los análisis. El propósito es determinar las especificidades de cada emisor.

Para el cumplimiento de dicho objetivo, se aplica un *análisis factorial de componentes principales* mediante el paquete estadístico BMDP. La *matriz de correlaciones* la forman cada uno de los 12 discursos que componen la *muestra* del estudio. Estos *discursos* actúan como *variables* y las 347 *palabras* como *casos.* De lo que se trata es de analizar las interrelaciones existentes entre ellos, en busca de semejanzas y especificidades.

La *matriz de correlaciones* entre las 12 variables se muestra en el Cuadro 10.2.

CUADRO 10.2. Matriz de correlaciones.

	1	2	3	4	5	6	7	8	9	10	11	12
1	1.00											
2	–.04	1.00										
3	–.10	–.06	1.00									
4	–.04	–.07	–.04	1.00								
5	–.12	–.05	–.03	–.03	1.00							
6	–.00	.05	–.07	–.08	–.10	1.00						
7	–.05	–.11	–.06	–.13	–.07	–.05	1.00					
8	–.05	–.10	–.09	–.16	–.12	–.05	.11	1.00				
9	–.14	.00	–.00	–.08	–.13	–.06	–.13	–.11	1.00			
10	–.15	–.10	–.03	.07	.06	–.19	–.25	–.30	.05	1.00		
11	–.04	–.09	–.15	–.10	–.10	–.13	–.20	–.20	–.12.	–.08	1.00	
12	–.12	–.06	–.08	–.17	–.10	–.12	–.17	–.25	–.13.	–.06	–.08	1.00

La *matriz factorial* de los 4 *factores* que resultan de la agrupación de las 12 variables, en función de la semejanza de los discursos, se rota por el procedimiento más habitual: *rotación ortogonal varimax*. Se busca la disposición de los *componentes* o *factores* de manera que cada uno de ellos tuviera saturaciones elevadas en pocas variables (en este caso, emisores de los discursos), y muy bajas en el resto.

El Cuadro 10.3 muestra las *cargas factoriales* de cada variable en los 4 *factores*. Todas las variables presentan, al menos, un *"factor loading"* superior a .25, en alguno de los *factores*. Como es usual en este tipo de análisis, los *factores* extraídos aportan de más a menos información: el primer *factor* explica más *varianza* (14%) que el segundo (11%), y éste más que el tercero (10%). La *varianza* explicada por el tercer *factor* es ligeramente superior a la explicada por el cuarto. Véase el Cuadro 10.3.

CUADRO 10.3. Matriz factorial rotada oblicuamente mediante el procedimiento varimax (*factor loadings* superiores a .25).

Emisor	Factor 1	Factor 2	Factor 3	Factor 4
GM	–	.517	–	–.394
VA	–	–	.570	–
FE	–	–.432	–	–
GS	–.268	–	–	–.681
JB	–	–	–.476	–
CH	–		.629	–
PA	.690	–	–	–
CC	.724	–	–	–
LS	–	–.562	–	–
RA	–.563	–.271	–.395	–
LF	–.253	.604	–	–
BA	–	–	–	–.726
Varianza	13,78%	10,82%	10,33%	10,06%

Para cada *factor* se presentan las formas (o *palabras)* positivas y negativas que lo componen, y sus emisores. En el Cuadro 10.4 se extractan algunas de las *palabras* correspondientes al *factor* 1. Este *factor* está caracterizado por los emisores CC, PA y RA, con unos *pesos* de .724, .690 y .563, respectivamente.

En los *valores positivos* se encuentran la mayoría de las *formas* que caracterizan tanto a CC como a PA. Destacan, especialmente, las correspondientes a los *ámbitos lingüísticos* ("lingua", "catalán2a", "catalán", "portugués2a", "ortografía", etc.), con resonancias de elementos discursivos del galleguismo de la época republicana.

El lado *negativo* del primer *factor* está fuertemente dominado por RA. En él se encuentra un léxico de corte más político, de claras referencias partidistas y electo-

rales ("PSOE", "CiU", "voto", "OTAN", "votar", "AP"); junto a un tímido eco del léxico sociológico y de estratificación social ("social", "clase", "sector", "sociedades"). En estos ámbitos el entrevistado RA mostraba una clara especialización.

CUADRO 10.4. Extracto de algunas de las formas positivas y negativas que componen el factor 1.

Factor 1				Emisores
		Formas positivas		
lingua	3.490	hispanico	1.540	CC .724
catalán2a	3.100	protagonismo	1.540	PA .690
catalán	2.750	mesmo	1.530	
portugués2a	2.710	galeguista	1.490	
ortografía	2.620	universal	1.470	
mentalidade	2.580	forza	1.460	
		Formas negativas		
pensar	-3.928	problema	-1.200	RA -.563
cousa	-2.078	caer	-1.183	
socia	-1.900	formulación	-1.122	
PSOE	-1.397	votar	-1.027	
diferente	-1.315	AP	-1.000	
OTAN	-1.209	traballar	-1.000	

A partir de las *formas positivas* y *negativas* que configuran cada *factor,* se da a cada uno de ellos las "etiquetas" o denominaciones siguientes:

- Factor 1: cultural-liberal / político-partidista
- Factor 2: sociológico-administración / político– estratificación social
- Factor 3: cultural-identitario / histórico-social
- Factor 4: estratégico-economía política / dispersión político-cultural

Los 4 *factores* se cruzan entre sí (el 1.º con el 2.º; el 1.º y el 3.º; el 1.º y el 4.º, y así), con la finalidad de comprobar la diferenciación léxica entre las distintas generaciones de nacionalistas entrevistados. Cada uno de los cruces se representan gráficamente mediante *planos de proyecciones* de las *palabras* que forman cada *factor*. De esta manera se obtienen las *estructuras léxicas* de las manifestaciones ideológicas.

El primer estadio de la investigación (el correspondiente al *análisis de contenido cuantitativo)* se da por concluido. Para un mayor conocimiento de esta primera etapa de investigación, así como el desarrollo del *análisis cualitativo,* remito a la publicación de esta excelente investigación de Julio Cabrera.

- Llovet, J. J. (1992): "El control de la prensa sobre la profesión médica: el caso de EL PAÍS", *Revista Española de Investigaciones Sociológicas,* n.º 59, pp. 261-285.

En esta segunda investigación se estudia el control social sobre la medicina a través de la prensa. Para ello se realiza un *análisis de contenido* centrado en el diario *El País*.

La *muestra* abarca las noticias, cartas de lectores, editoriales, etc., publicados en el lustro 1985-1989, referidos a negligencia y conductas médicas incorrectas o lesivas. Se descartaron, por el contrario, las informaciones que aludían a hechos o procesos de este carácter ocurridos fuera de España; al igual que los derivados de la atención psiquiátrica (salud mental), por ser ésta un área muy específica.

En primer lugar se lleva a cabo un rápido peinado visual de los titulares (antetítulos y títulos), aparecidos en las secciones del diario en las que corrientemente pueden leerse comentarios o información sobre el tema. Pocos meses del año quedan en blanco, sin esta clase de información.

En el Cuadro 10.5 se exponen las *distribuciones porcentuales* de las unidades de redacción de *El País*. Éstas se refieren a negligencias y actuaciones incorrectas, o lesivas, de médicos, según la subsección del diario y género periodístico (del 1 de enero de 1985 al 31 de diciembre de 1989).

CUADRO 10.5. Distribuciones porcentuales de unidades de redacción de *El País*.

Frecuencia porcentual	
Subsección del diario	
"Opinión"/"Cartas al director"	22,9
"Sociedad"/"Tribunales"	22,2
"Sociedad"/"Sanidad"	21,5
"Sociedad" (sin especificación)	15,6
"España"	7,3
"Madrid"	5,2
"Opinión"/"Editorial"	2,8
"Domingo"	1,0
"Sociedad"/"Ciencia"	0,3
"Contra-tapa"	0,3
"Opinión"/"El *Ombudsman*"	0,3
"Opinión"/"Fe de errores"	0,3
Total	100
	(288)
Género periodístico	
Noticia	70,1
Carta de lector	22,9
Editorial	2,8
Reportaje, informe, reseña	2,8
Artículo o columna firmada	1,0
Errata	0,3
Total	100
	(288)

También se calculan las *frecuencias porcentuales* de las noticias, según el papel social o institucional de los enunciadores: jueces o magistrados, médicos denunciados, familiares o amigos de los pacientes perjudicados, pacientes perjudicados, abogados de los pacientes, colegios de médicos, etc.

De toda la información extraída del periódico *El País,* se deduce la contribución de la prensa como factor preponderante en la apertura, esclarecimiento y sensibilización de los ciudadanos, y pacientes, hacia la problemática sanitaria.

Como puede verse, el uso que del *análisis de contenido* se ha hecho en esta segunda investigación es bastante más simple que el efectuado en la primera investigación. Se limita a la descripción de frecuencias y comentario de extractos de textos que aparecieron en prensa durante el período analizado. No obstante, el principal interés en la exposición de esta segunda investigación es ilustrar un modo de hacer *análisis de contenido,* también usual en la investigación social.

10.5. La aplicabilidad del análisis de contenido y los límites a la inferencia

La aplicación del *análisis de contenido* se ha vinculado, tradicionalmente, al estudio de la comunicación. En 1969 Holsti sintetiza su *aplicabilidad* en la resolución de los siguientes interrogantes de la comunicación:

a) Qué se dice:

- Describir tendencias y diferencias en el "contenido" de la comunicación.
- Relacionar las características de las fuentes con los mensajes que éstas generan.
- Evaluación del "contenido".
- Extraer los patrones culturales presentes en el mensaje.

b) Cómo se dice:

- Comprobar si en el mensaje se hace uso de técnicas persuasivas.
- Analizar el estilo del texto.

c) A quién:

- Qué peculiaridades presentan aquéllos a quienes está destinado el mensaje.

d) Por qué:

- Identificar las intenciones y otras características de los emisores del mensaje.

e) Quién lo dice:

- Determinar la autoría del emisor.

f) Con qué efectos:

- Analizar el flujo de la comunicación.

Sin embargo, la aplicación que del *análisis de contenido* se hace desde la vertiente metodológica cuantitativa se halla condicionada por *limitaciones* importantes:

a) Aunque la *cuantificación* ayude a la sistematización y estandarización del análisis textual, no siempre se profundiza en los *contenidos no manifiestos* de los mensajes.
b) La interpretación de las *unidades de codificación* no ha de restringirse al mero recuento de *frecuencias* de aparición. Por el contrario, exige su previa contextualización. Precisamente, porque unas mismas palabras, en contextos diferentes, pueden implicar significados distintos.

Estas limitaciones pueden solventarse complementando el *análisis de contenido "categorial"* con el *"contextual"*, el *"cuantitativo"* con el *"cualitativo"* (este último más centrado en la indagación de las *estructuras semánticas* y *semiológicas* de los documentos). Como bien reconocen Ruiz Olabuenaga e Ispizua (1989: 185), "los mejores análisis de contenido son aquellos que utilizan la técnica de la *triangulación*". Ello permite una mejor captación de su pleno significado, la definición de la situación y del punto de vista del emisor.

10.5.1. Cuestiones de validez y de fiabilidad

La *inferencia* en el *análisis de contenido* está condicionada, como en cualquier procedimiento analítico, a la *validez* y *fiabilidad* de la *medición* aplicada. En este caso, la síntesis hecha de los "contenidos" de los documentos que se analizan.

La *validez* indica la *adecuación* del sistema de clasificación seguido en la medición de los *conceptos* teóricos; mientras que la *fiabilidad* expresa la *consistencia* del procedimiento de medición.

• Respecto a la *fiabilidad,* han de arbitrarse unas reglas de *codificación* que permitan que distintos codificadores alcancen, independientemente unos de otros, los mismos resultados. Esto es más fácilmente alcanzable cuando el *análisis de contenido* se realiza mediante ordenador. Una vez que las *reglas de codificación* se han definido, éstas se aplican automáticamente, y de igual manera en cada uno de los casos analizados.

Krippendorff (1980, 1990) diferencia tres tipos de *fiabilidad* pertinentes en el *análisis de contenido:*

a) *Estabilidad:* grado en el que varían los resultados de la clasificación a lo largo del tiempo.

Deberían alcanzarse los mismos resultados en la clasificación de las *unidades de codificación,* siempre que se emplee la misma agrupación de *categorías*. Ello exige la eliminación de cualquier ambigüedad en la clasificación de las *unidades* en categorías.

b) *Reproducibilidad:* grado al que se alcanzan los mismos resultados, cuando un mismo texto se codifica por varias personas.

En este caso, la *infiabilidad* se atribuye, sobre todo, a arbitrariedades en la *codificación*. Estas pueden deberse a ambigüedades en las instrucciones de codificación. Pero, también a errores de grabación o a diferencias cognitivas entre las personas que realizan la *codificación*.

c) *Adecuación:* grado en el que la clasificación del documento se corresponde con un estándar o norma.

Como rara vez se establecen *codificaciones estándar,* este último criterio de *fiabilidad* apenas se aplica.

Para la cuantificación de la *fiabilidad* se han propuesto algunos *coeficientes*. Sánchez Carrión (1985) destaca los dos siguientes:

a) La *razón de codificaciones de acuerdo,* entre el total de codificaciones:

$$C.R.=\frac{2M}{N_1+N_2}$$

donde: "M" corresponde al número de codificaciones en las que coinciden los 2 codificadores.
"N_1" y "N_2", el número de codificaciones efectuadas.

b) El *índice de fiabilidad* (p_i) de Scott:

$$p_1=\frac{proporción\ \ de\ \ acuerdo\ \ observado-proporción\ \ de\ \ acuerdo\ \ esperado}{1-proporción\ \ de\ \ acuerdo\ \ esperado}$$

Este índice oscila entre .00 y 1.00. El valor 1.00 representa el acuerdo total entre los codificadores; mientras que .00, el desacuerdo.

La *validez,* en cambio, denota correspondencia entre los *constructos teóricos* y las *variables empíricas* empleadas. Esto exige, igualmente, la eliminación de cualquier ambigüedad en la clasificación de las *unidades de codificación* en *categorías*.

En concreto, se diferencian varios tipos de *validez* en el *análisis de contenido* (López Aranguren, 1989; Krippendorff, 1990; Weber, 1994):

a) *Validez pragmática* o de utilidad de los resultados del análisis en el estudio del tema que se investiga.
b) *Validez de constructo:* correspondencia entre la medida aplicada con alguna otra utilizada del mismo *constructo.*
c) *Validez de contenido:* grado en el que los *indicadores* cubren la variedad de significados incluidos en el *concepto.*

d) Validez predictiva: grado de acuerdo/correspondencia entre las predicciones obtenidas y los hechos observados.

Las *inferencias* hechas a partir del análisis de la información han de "mostrar un alto grado de acuerdo con los estados, atributos, sucesos o propiedades del contexto de los datos al que esas inferencias se refieren, y un alto grado de desacuerdo con las características contextuales que esas diferencias excluyen lógicamente" (Krippendorff, 1990: 232).

e) Validez semántica: grado en el que las *unidades de codificación* clasificadas juntas poseen connotaciones similares.

En el *análisis de contenido* "se logra una alta validez semántica cuando la semántica del lenguaje de los datos se corresponde con la fuente, el receptor o cualquier otro contexto respecto al que se examinan dichos datos" (Krippendorff, 1990: 231).

Los límites a la *inferencia* se hallan, en conclusión, en la ambigüedad de los significados de las *palabras* y de las *reglas de codificación*. Por lo que se recomienda su explícita delimitación.

Lecturas complementarias

Bardin, L. (1986): *Análisis de contenido*. Madrid, Akal.

Behar, J. (1993): "Aproximación al análisis textual informatizado". *Anuario de Psicología,* n.º 59, pp. 61-78.

Clemente Díaz, M. (1992): "El análisis de contenido". En Clemente, M. (comp.): *Psicología Social: métodos y técnicas de investigación*. Madrid, Eudema, pp. 169-207.

Krippendorff, K. (1990): *Metodología del análisis de contenido: teoría y práctica.* Barcelona, Paidós.

López-Aranguren, E. (1994): "El análisis de contenido". En García Ferrando, M. *et al.* (comp.): *El Análisis de la Realidad Social.* Madrid, Alianza, pp. 461-492.

Mochmann, E. (1985): "Análisis de contenido mediante ordenador aplicado a las ciencias sociales". *Revista Internacional de Sociología,* vol. 43 (1), pp. 11-44.

Sánchez Carrión, J. J. (1985): "Técnicas de análisis de los textos mediante codificación manual". *Revista Internacional de Sociología,* vol. 43 (1), pp. 89-118.

Weber, R. P. (1994): "Basic content analysis". En Lewis-Beck, M. S. (comp.) *Research Practice.* California, Sage, pp. 251-338.

EJERCICIOS PROPUESTOS

1. Describa una investigación en la que se haya hecho un análisis de contenido cuantitativo.
2. Analice el contenido de los editoriales del próximo fin de semana que aparezcan en tres periódicos.
3. ¿Qué peculiaridades presentan los criterios de validez y de fiabilidad en el análisis de contenido? Señale semejanzas y diferencias con los enunciados en el Capítulo 4.
4. Un grupo de investigadores desea analizar la influencia de la prensa escrita en el desarrollo de una campaña electoral. Para ello deciden analizar los contenidos de periódicos publicados durante la última campaña electoral. Diseñe la investigación: especifique cómo se seleccionaría la muestra del estudio, cuáles serían las unidades de codificación y de contexto, y trace el plan de análisis.

11

EL INFORME DE LA INVESTIGACIÓN

La investigación sociológica culmina con la redacción del *informe*. En él cristalizan cada uno de los estadios que forman el proceso investigador: desde la formulación inicial del problema y la delimitación de los objetivos, hasta la interpretación de los resultados de la investigación. Sin esta exposición pormenorizada del estudio (de su desarrollo y del alcance de los hallazgos) no podría darse por concluida la investigación. El investigador estaría más lejos de poder cumplir uno de los requisitos fundamentales de su trabajo: la comunicación pública en general y a la comunidad científica en particular, para su conocimiento y valoración. Como Caplow (1977: 15) afirma:

> "El objetivo oficial del investigador es siempre el de contribuir al enriquecimiento del conocimiento científico. Se sobrentiende que aquél publicará los resultados obtenidos sin restricciones ni deformaciones, y que proporcionará suficientes informaciones sobre sus métodos a fin de que otros investigadores puedan reproducir este estudio si así lo desean."

El *informe* adquiere tal relevancia en el esfuerzo investigador, que Caplow (1977) llega incluso a negar la consideración de "investigación sociológica" a aquélla cuyos resultados no sean publicados.

A continuación se exponen algunas consideraciones genéricas, a tomar como referencia, en la elaboración de un *informe* de investigación.

11.1. La organización del informe: contenidos básicos

El *informe* cumple la función de presentación, de comunicación, de un esfuerzo investigador. Razón por la cual ha de referirse explícitamente a las distintas fases que

componen la investigación, desde su concepción inicial hasta su conclusión. Se recomienda que esta exposición sea objetiva, transparente, y lo más fidedigna posible respecto a los contenidos de la investigación. En palabras de Alvira (1995: 246-247):

> "El informe tiene que dar cuenta del cómo, cuándo y con qué resultados se han cumplido los objetivos del proyecto. Un informe debe explicar los procesos seguidos, las divergencias o cambios en lo proyectado (si los hubiere) y la confirmación o refutación de las hipótesis de partida que dieron lugar al proyecto de investigación."

Representa, en suma, la *síntesis,* la culminación del ejercicio investigador, en forma de documento para su lectura y evaluación. Este documento puede adoptar distintas formas en conformidad con las características del público al que principalmente se dirige (profano, experto), del medio empleado en su divulgación (monografía, revista, comunicación de congreso, informe técnico), y del ingenio y estilo comunicativo del investigador, fundamentalmente.

Dependiendo de cómo se describa la investigación (genérica o pormenorizadamente), de qué aspectos se antepongan (los metodológico-técnicos al comentario de resultados), y del grado de tecnicismo que se dé a la redacción de los contenidos de la investigación, se diferencian varios tipos de informes. Desde los *técnicos* (dirigidos a una audiencia interesada en el conocimiento de todo el desarrollo metodológico de la investigación) a los *divulgativos* (centrados en el comentario de los resultados de la investigación, sin tecnicismos, en un lenguaje accesible al público general).

No obstante, en la mayoría de los informes de investigación puede encontrarse una *estructura básica común,* que adquirirá distinto desarrollo en función de las características del *informe.* Esta *estructura* está integrada por seis grandes apartados:

a) Introducción.
b) Metodología.
c) Resultados.
d) Discusión.
e) Conclusión.
f) Bibliografía.

• *Introducción*

El *informe* suele comenzar con una sección introductoria, que ofrece una visión global de la investigación y su relevancia para el conocimiento del problema que se analiza. El investigador define y contextualiza el *problema* de investigación, especificando los *objetivos* e *hipótesis* esenciales en la indagación. Pero también resalta la trascendencia del esfuerzo investigador realizado respecto a indagaciones anteriores.

En su justificación de la investigación, el investigador destaca el interés de su estudio y por qué constituye éste un avance en el conocimiento del problema que ana-

liza. Para ello contextualiza la investigación en su área de conocimiento. El propósito no es mostrar un catálogo de la literatura especializada en el tema, sino señalar cuál es el estado actual de la cuestión: qué cuestiones se han resuelto y sobre cuáles existen aún controversias o una falta de estudios.

Dicha argumentación se acompaña de un número limitado de referencias (de estudios de notorio interés), seleccionadas para marcar la trayectoria investigadora precedente.

• *Metodología*

Bien después de la "introducción", bien al final del *informe,* en un "apéndice metodológico", se describen, con mayor o menor detalle (dependiendo del *informe),* los aspectos relacionados con el *diseño de la investigación.* Concretamente, se expone qué *estrategia* se siguió en la indagación del problema de estudio; se datallan las características técnicas de la *muestra;* y se ofrecen comentarios, que se juzguen pertinentes, sobre las *técnicas de obtención* y de *análisis* de datos empleadas. Además, se comentan las incidencias (si las hubiera) en la materialización del *diseño* proyectado, señalándose las actuaciones para su resolución.

El propósito fundamental de este apartado –como indica Gilbert (1993: 338)– reside en "proporcionar el detalle suficiente para que otro investigador pueda repetir la investigación". Por esta razón han de explicitarse cada uno de los pasos seguidos en la ejecución del *proyecto* de investigación: la operacionalización de los *conceptos teóricos* fundamentales, la selección de la *muestra,* la fecha del *trabajo de campo,* el porcentaje de respuestas, junto con la descripción de las *técnicas* utilizadas para la *recogida* (cuestionario, guión de entrevista...) y *análisis* de la información.

Además de describir las singularidades del *diseño* de la investigación, es importante que se indiquen las posibles limitaciones de la información obtenida: su *validez* y *fiabilidad.*

Por *ejemplo,* Fowler (1988: 143) indica, en el caso de una *encuesta,* la siguiente información a incluir en el *informe* de la investigación:

a) El *marco de muestreo* y la estimación del porcentaje de la *población* que tuvo oportunidades de selección. A ser posible, también señalar si la *población* excluida de la *muestra* difiere de aquella que finalmente participa en el estudio.
b) El *procedimiento de muestreo* seguido en la selección muestral.
c) *Resultados del campo:* porcentaje de respuestas y razones principales para la "no respuesta".
d) Descripción de *quién llevó a cabo la entrevista:* si se contrataron entrevistadores y supervisión efectuada.
e) Breve descripción del *diseño del cuestionario,* incluyendo su precomprobación *(pretest).*
f) La *redacción exacta de las preguntas,* si éstas no están reproducidas en el texto.

El *informe* debería incluir el *cuestionario* íntegro, al menos en un anexo al final del documento. El conocimiento de la redacción exacta de las preguntas y orden de las mismas, sin duda, ayudará al lector en la valoración de la información publicada en el *informe.*

El *cuestionario* suele figurar al final del documento, acompañado de las frecuencias obtenidas en cada opción de respuesta de las preguntas.

g) La calidad de los procedimientos de comprobación y control durante la *codificación, grabación* y *análisis* de los datos.

h) Un breve comentario sobre los posibles *errores* en la *encuesta.*

• *Resultados*

Descrita la metodología, se pasará en el *informe* a la relación y comentario de los "hallazgos" de la investigación. La atención se pone, sobre todo, en los *resultados* más relevantes respecto a los *objetivos* del estudio.

La exposición de los *resultados* ha de caracterizarse por su claridad, precisión y objetividad. El texto puede acompañarse con *gráficos* y *tablas* que faciliten la exposición y la lectura de los resultados numéricos del estudio.

• *Discusión*

Los *resultados* se interpretan relacionándolos con las *hipótesis* iniciales de la investigación. Se discute si éstas son corroboradas por la información reunida.

En esta *discusión* también se comparan los *resultados* de la investigación con los de estudios (si los hubiese) anteriores, además de precisar la relevancia y los límites a su generalización.

De toda esta reflexión se deducirá la respuesta de algunas de las cuestiones planteadas en la *introducción* del *informe.* Otras, sin embargo, quedarán sin respuesta, poniendo de manifiesto las limitaciones del estudio. Por lo que es común que este apartado concluya con la propuesta de futuras líneas de investigación.

• *Conclusión*

De todas los apartado del *informe,* éste es –de acuerdo con Gilbert (1993: 340)– el "más difícil de escribir porque debería resumir cuáles fueron los principales hallazgos de la investigación de una forma sucinta e interesante. Esto es, porque muchos lectores mirarán primero el final para ver si tiene algo interesante que decir, antes de comenzar por el principio".

Se relatan los hallazgos más significativos del estudio, dejando la puerta abierta a futuras indagaciones que resuelvan cuestiones discutidas o sin respuesta en la investigación descrita.

• *Bibliografía*

A lo largo del *informe* se habrá hecho referencia a otros estudios de interés, sobre el mismo *problema* o problemas colaterales. Estas referencias figuran en el texto como *citas*, con el primer apellido del autor, seguido de la fecha de publicación de la obra, todo escrito entre paréntesis, generalmente. Si de la obra se ha extraído literalmente un fragmento, a la *cita* se adjunta el número de las páginas en que dicho fragmento se halla. Esto permite que el lector pueda localizar la *cita* en su contexto original. Pero, para ello debe previamente conocer la referencia completa de la obra.

Al final del *informe* figuran las *referencias bibliográficas* completas de los textos consultados en la investigación, normalmente ordenadas alfabéticamente por autor.

En las *referencias* de *libros* se escribe el apellido principal del autor y las iniciales del nombre, la fecha de publicación de la obra (entre paréntesis), el título del libro (generalmente en cursiva), la ciudad donde se publica y el nombre de la editorial. Para las *revistas,* el título del artículo aparece entre comillas, acompañado del nombre de la revista (en cursiva), el número o volumen de la misma, y las páginas correspondientes al artículo.

En la *bibliografía* han de figurar, únicamente, las referencias exactas de aquellos estudios a los cuales se ha hecho mención en el *informe*. Por tanto, se trata de incluir sólo las obras realmente consultadas en la realización y redacción del estudio.

A este esquema genérico de *informe* a veces se añade un breve *resumen,* que contiene la información más notoria de la investigación. Este *resumen* puede figurar al principio o al final del *informe*.

Si el *informe* presenta la forma de *artículo* de revista, por ejemplo, el *resumen* aparece al principio del artículo: después del *título* y antes de la *introducción*. Su extensión es muy reducida. En un máximo de 10 líneas, suele resumirse el contenido principal del artículo. Este breve *resumen* cumple la función de ofrecer una visión sintética y global de la investigación, con el propósito de captar la atención del lector y despertar su interés por su lectura.

También es frecuente dedicar unas líneas, al comienzo del *informe,* a los *agradecimientos* o mención de las personas e instituciones que contribuyeron en la realización de la investigación: entidades que la financiaron o colaboraron de algún modo en la materialización del estudio; y particulares que ayudaron al equipo investigador durante el curso de la investigación o en su conclusión (en la lectura del borrador del *informe,* por ejemplo).

La *estructura del informe,* aquí expuesta, puede experimentar variaciones según la especificación y complejidad que el investigador dé a su *informe*. Sobre todo dependerá de la formación e interés del lector al que el *informe* se dirige principalmente.

> "Un informe escrito como un documento privado para el departamento o empresa que financia la investigación enfatizará cosas diferentes, quizás dando menos espacio a la metodología, que uno dirigido a expertos. Ambos serán más formales y técnicos que un informe dirigido al público general" (Moser y Kalton, 1989: 467).

11.2. La redacción del informe

La redacción del *informe* supone el esfuerzo de plasmar y sintetizar buena parte de las tareas llevadas a cabo durante la investigación. La escritura de lo hecho y su justificación puede resultar para algunos sencilla, mientras que para otros se convierte en una labor tediosa. Dependerá de la práctica del investigador, de su capacidad creativa, así como de su habilidad en las artes de la comunicación escrita.

En este último apartado lo único que se pretende es señalar algunas sugerencias que ayuden a afrontar la redacción de un *informe* de investigación.

Para la *redacción del informe* es muy importante que el investigador identifique, previamente, las características e intereses de su *audiencia.* Los rasgos de ésta habrán de tenerse en cuenta a la hora de dar forma al *informe.* Reléase la cita de Moser y Kalton, reproducida más arriba.

En concreto, de la *audiencia* interesa conocer:

a) Su grado de conocimiento sobre el tema expuesto.
b) El lenguaje en que se expresa.
c) Las cuestiones que más le interesan.

Este conjunto de aspectos llevará al investigador a escribir de una forma más técnica o, por el contrario, más coloquial; a exponer los contenidos de la investigación de una forma más detallada o más genérica; a poner mayor énfasis en la metodología o en los resultados.

> "Si se escribe para científicos sociales, pueden hacerse determinados supuestos sobre su conocimiento previo, y quizás puedan resumirse ciertos puntos más que explicarse en detalle. Igualmente puede emplearse apropiadamente un lenguaje más técnico que el apropiado para una audiencia general. Términos y supuestos aceptables para los colegas próximos pueden sólo confundir a otros científicos" (Babbie, 1992: A9).

Si bien, siempre se ha de procurar facilitar la lectura del *informe.* De ahí la usual recomendación de traducir los *tecnicismos* a un lenguaje conocido por el lector potencial de la investigación. La *redacción* ha de ser lo más clara y accesible posible.

> "No importa para quién estás escribiendo, mantén tus frases cortas y sencillas, con preferencia a los términos técnicos. Las construcciones complejas sólo confunden al lector (y podrían llevar a la sospecha de que tú, el escritor, no estás pensando con claridad tampoco)" (Gilbert, 1993: 334).

Para la *redacción del informe* también es importante la identificación del *medio* que va a utilizarse para la difusión de la investigación. Si va a publicarse dentro de una colección de monografías o en una revista científica, resulta conveniente la consulta previa de algunos de los números editados con anterioridad. De su lectura, el investiga-

dor podrá conocer qué formato se espera que cumpla su *informe:* su estructura, extensión, y redacción de los contenidos.

Identificada la *audiencia* y el *medio,* el investigador puede comenzar la *redacción del informe.* Para ello tan importante o más que lo anterior es la previa elaboración de un *esquema* o *índice,* que reúna los principales puntos a cubrir en el *informe.*

Este *esquema* suele ajustarse a la estructura general expuesta en el apartado anterior, en correspondencia con las distintas fases del *proyecto de investigación.* Con él se pretende organizar la información de manera que se facilite su escritura.

El investigador puede comenzar por el primer apartado y seguir el orden del *índice* hasta el final. Pero, puede también actuar siguiendo otros órdenes y cubrir a saltos el *esquema* inicial.

Una práctica usual es dejar la redacción de la *introducción* y la *conclusión* para el final. Como ambos apartados ofrecen una visión conjunta de la investigación, lo más práctico e inteligente es aplazar su redacción para el final. Si se opta por lo opuesto, la redacción de la *introducción* puede verse alterada tras la escritura del *informe* completo.

También, conviene tener presente que la *introducción* y la *conclusión* suelen ubicarse al principio y al final del *informe,* respectivamente. Posiciones éstas consideradas estratégicas para captar la atención del lector. Por lo que ha de ponerse especial cuidado en su redacción.

Concluida la redacción del *informe,* siempre es aconsejable que otros profesionales lean el documento antes de su presentación o edición. Sus sugerencias siempre suponen mejoras.

La edición definitiva del *informe* va, con frecuencia, precedida de varios borradores y el empleo de muchas horas de trabajo, pues "la redacción de la investigación siempre implica mucha reescritura y refinamiento" (Gilbert, 1993: 334).

Lecturas complementarias

Alexander, C. (1986): *Writing for social scientists: how to start and finish your thesis, books or article.* Chicago, Chicago University Press.

Alvira, F. (1995): "Formas de presentación y representación gráfica de los datos. Preparación del informe". En MAP (ed.): *Sociología. Temas de materias sociales,* vol. 2, pp. 245-275.

Calvert, P. (1991): "Writing skills". En Allan, G. y C. Skinner (comp.) *Handbook for research students in the social sciences.* London, The Falmer Press, pp. 96-106.

Cooper, H. M. (1984): *The integrative research review: a systematic approach.* Beverly Hills, Sage.

Eco, U. (1982): *Cómo se hace una tesis.* Barcelona, Gedisa.

Gilbert, N. (1993): "Writing about social research". En Gilbert N. (comp.) *Researching social life.* Londres, Sage, pp.327-344.

Richlin, J. y Strenski E. (1991): *A guide to writing sociology papers.* New York, Saint Martin.

Sierra Bravo, R. (1986): *Tesis doctorales y trabajos de investigación.* Madrid, Paraninfo

Ejercicios propuestos

1. Elija dos investigaciones, una publicada en una revista científica y la otra como monografía. Compare la presentación que en ambas se hace de la investigación: la estructura y redacción del informe.
2. Elabore un proyecto de investigación sobre un tema de su interés. En él incluya todos los componentes esenciales de un proyecto y justifique el diseño de su investigación: por qué ha elegido unas estrategias y técnicas con preferencia a otras.